WHO Life Expectancy — Modeling & Variable Selection

Dominion Samuel

August 26, 2025

Introduction

This project stems from my deep interest in improving health outcomes in developing countries. I was motivated to explore the key drivers of low life expectancy to better understand which areas could benefit most from targeted interventions. Using global health data, this analysis investigates factors associated with low life expectancy (defined as under 65 years) and employs logistic regression (binary outcome), stepwise selection (AIC and BIC), and LASSO regularization to identify the most influential predictors and assess model performance.

I aim to:

1. Transform life expectancy into a binary outcome.
2. Build predictive models using stepwise selection and LASSO.
3. Compare selected variables and evaluate the final model using accuracy, confusion matrices, and ROC/AUC.
4. Interpret my findings and their implications in potential decisions.

Data Preparation

```r
library(tidyverse)
library(broom)
library(MASS)     # stepAIC
library(glmnet)   # LASSO
library(caret)    # train/test split + confusion matrices
library(pROC)     # AUC/ROC
```

```r
# Load cleaned life expectancy dataset
life <- read_csv("../data/life_clean.csv")

# Create binary outcome: 1 if life expectancy < 65, else 0
life <- life %>%
  mutate(
    low_lifeexp = ifelse(life_expectancy < 65, 1, 0),
    low_lifeexp = factor(low_lifeexp, levels = c(0,1))
  ) %>%
  dplyr::select(-life_expectancy)
```

```r
# Overview of dataset
glimpse(life)
```

```
## Rows: 1,407
## Columns: 15
## $ adult_mortality     <dbl> 263, 271, 268, 272, 275, 279, 281, 287, 295, 295, ~
## $ infant_deaths       <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84, 85, 87, 87~
## $ alcohol             <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.03, 0.~
## $ hepatitis_b         <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64, 66, 67, 65~
## $ measles             <dbl> 1154, 492, 430, 2787, 3013, 1989, 2861, 1599, 1141~
## $ bmi                 <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16.2, 15.7, 15~
## $ polio               <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41, ~
## $ total_expenditure   <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.42, 8.33, 6.~
## $ diphtheria          <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41,~
## $ hiv_aids            <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, ~
## $ gdp                 <dbl> 584.25921, 612.69651, 631.74498, 669.95900, 63.537~
## $ population          <dbl> 33736494, 327582, 31731688, 3696958, 2978599, 2883~
## $ thinness_1_19_years <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18.6, 18.8, 19~
## $ schooling           <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8.7, 8.4, 8.1~
## $ low_lifeexp         <fct> 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,~
```

```r
# Check balance of outcome variable
table(life$low_lifeexp)
```

```
##
##   0   1
## 968 439
```

The outcome is imbalanced (968 vs 439), but not extreme

```r
# Split data into 70% train, 30% test for later testing
set.seed(42)
train_idx <- sample(nrow(life), 0.7*nrow(life))
train <- life[train_idx, ]
test  <- life[-train_idx, ]
```

Logistic Regression

```r
# Base model with all predictors
base_formula <- as.formula("low_lifeexp ~ .")

m_base <- glm(base_formula, data = train, family = binomial())
```

We get warnings about fitted probabilities numerically 0 or 1, likely due to separation or extreme predictors. This is common with rare or perfectly separable events and should not affect analysis strongly.

Stepwise Selection

```r
m_step_aic <- stepAIC(m_base, direction = "both", trace = FALSE)

n_train <- nrow(train)
m_step_bic <- stepAIC(m_base, direction = "both", k = log(n_train), trace = FALSE)

list(
  AIC_selected = formula(m_step_aic),
  BIC_selected = formula(m_step_bic)
)
```

```
## $AIC_selected
## low_lifeexp ~ adult_mortality + diphtheria + hiv_aids + gdp +
##     schooling
##
## $BIC_selected
## low_lifeexp ~ adult_mortality + hiv_aids + gdp + schooling
```

AIC selected: adult_mortality + diphtheria + hiv_aids + gdp + schooling

BIC selected: adult_mortality + hiv_aids + gdp + schooling

BIC is stricter, penalizing model complexity more heavily.

LASSO Selection

```r
# Model matrices for glmnet
x_tr <- model.matrix(base_formula, data = train)[,-1]
y_tr <- as.numeric(train$low_lifeexp) - 0

x_te <- model.matrix(base_formula, data = test)[,-1]
y_te <- as.numeric(test$low_lifeexp) - 0


# 10-fold CV for LASSO
cvfit <- cv.glmnet(
  x = x_tr, y = y_tr,
  alpha = 1,            # LASSO
  family = "binomial",
  nfolds = 10
)

# Extract selected coefficients
coef_lasso <- coef(cvfit, s = "lambda.1se")
lasso_tbl <- tibble(
  feature = rownames(coef_lasso),
  coef = as.numeric(coef_lasso)
) %>%
  filter(feature != "(Intercept)", abs(coef) > 1e-8) %>%
  arrange(desc(abs(coef)))

lasso_tbl
```

```
## # A tibble: 7 x 2
##   feature               coef
##   <chr>                <dbl>
## 1 schooling         -0.544
## 2 hiv_aids           0.411
## 3 total_expenditure  0.0200
## 4 bmi               -0.0107
## 5 adult_mortality    0.00961
## 6 diphtheria        -0.00509
## 7 polio             -0.00209
```

LASSO shrinks many coefficients to zero, highlighting the most important predictors for low life expectancy.

```
# Fit final logistic regression using LASSO-selected variables
selected_vars <- lasso_tbl$feature
formula_sel <- as.formula(paste("low_lifeexp ~", paste(selected_vars, collapse = " + ")))

model_final <- glm(formula_sel, data = train, family = binomial())
summary(model_final)
```

```
##
## Call:
## glm(formula = formula_sel, family = binomial(), data = train)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.556032   0.921728   3.858 0.000114 ***
## schooling          -0.679371   0.084168  -8.072 6.94e-16 ***
## hiv_aids            0.998913   0.140152   7.127 1.02e-12 ***
## total_expenditure   0.107420   0.076184   1.410 0.158534
## bmi                -0.013709   0.009457  -1.450 0.147155
## adult_mortality     0.013006   0.001813   7.175 7.23e-13 ***
## diphtheria         -0.008275   0.006671  -1.241 0.214787
## polio              -0.002013   0.006468  -0.311 0.755678
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1223.09  on 983  degrees of freedom
## Residual deviance:  371.51  on 976  degrees of freedom
## AIC: 387.51
##
## Number of Fisher Scoring iterations: 8
```

Significant predictors: schooling, hiv_aids, adult_mortality

Other variables (total_expenditure, bmi, diphtheria, polio) are not significant individually.

Direction: Higher schooling decreases odds of low life expectancy; higher HIV/AIDS prevalence and adult mortality increase odds.

Model Evaluation

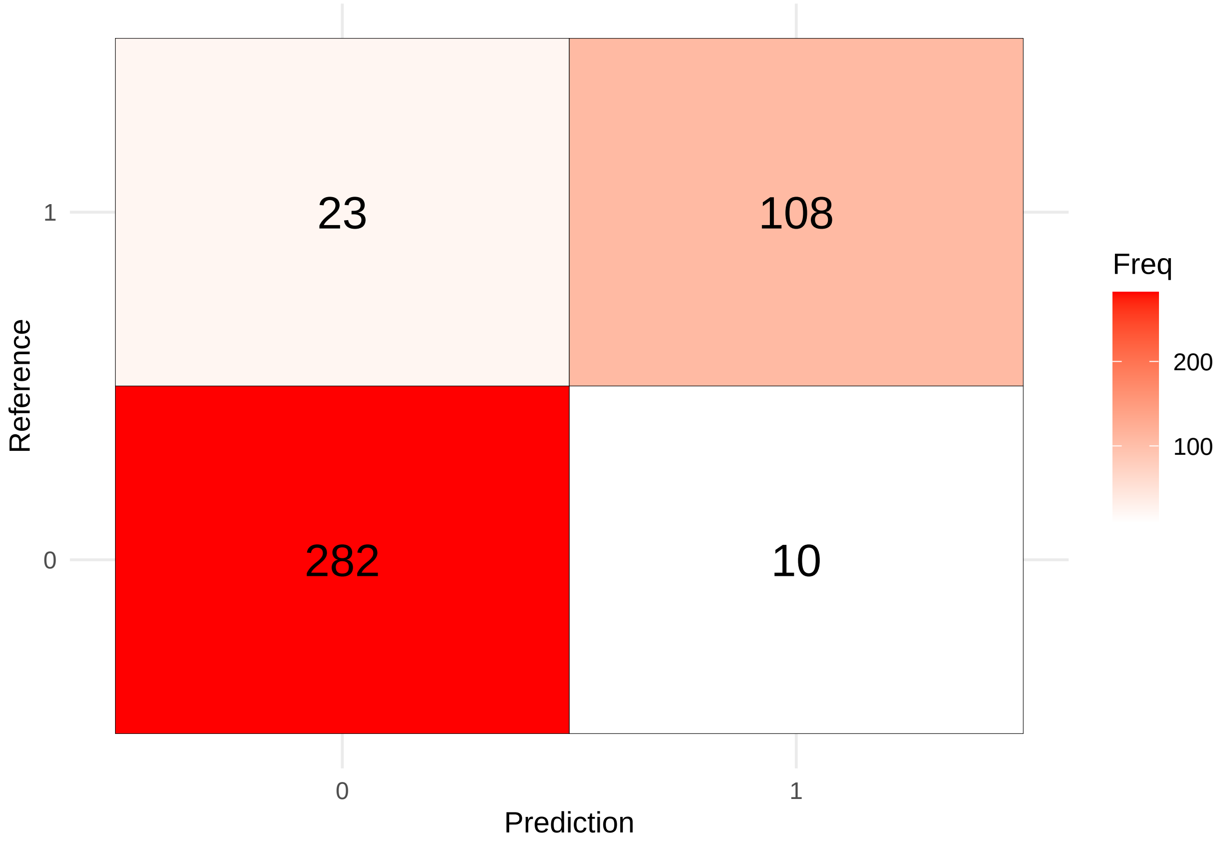
```r
# Predictions on test set
pred_prob <- predict(model_final, test, type = "response")
pred_class <- factor(ifelse(pred_prob >= 0.5, 1, 0), levels = c(0,1))

# Confusion matrix
conf_mat<- confusionMatrix(pred_class, test$low_lifeexp)
conf_mat
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 282  23
##          1  10 108
##
##                Accuracy : 0.922
##                  95% CI : (0.8922, 0.9457)
##     No Information Rate : 0.6903
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8124
##
##  Mcnemar’s Test P-Value : 0.03671
##
##             Sensitivity : 0.9658
##             Specificity : 0.8244
##          Pos Pred Value : 0.9246
##          Neg Pred Value : 0.9153
##              Prevalence : 0.6903
##          Detection Rate : 0.6667
##    Detection Prevalence : 0.7210
##       Balanced Accuracy : 0.8951
##
##        ’Positive’ Class : 0
##
```

```r
cm_table <- as.table(conf_mat$table)
cm_df <- as.data.frame(cm_table)

ggplot(cm_df, aes(Prediction, Reference, fill = Freq)) +
  geom_tile(color = "black") +
  geom_text(aes(label = Freq), size = 6) +
  scale_fill_gradient(low = "white", high = "red") +
  theme_minimal()
```



Accuracy: 0.922

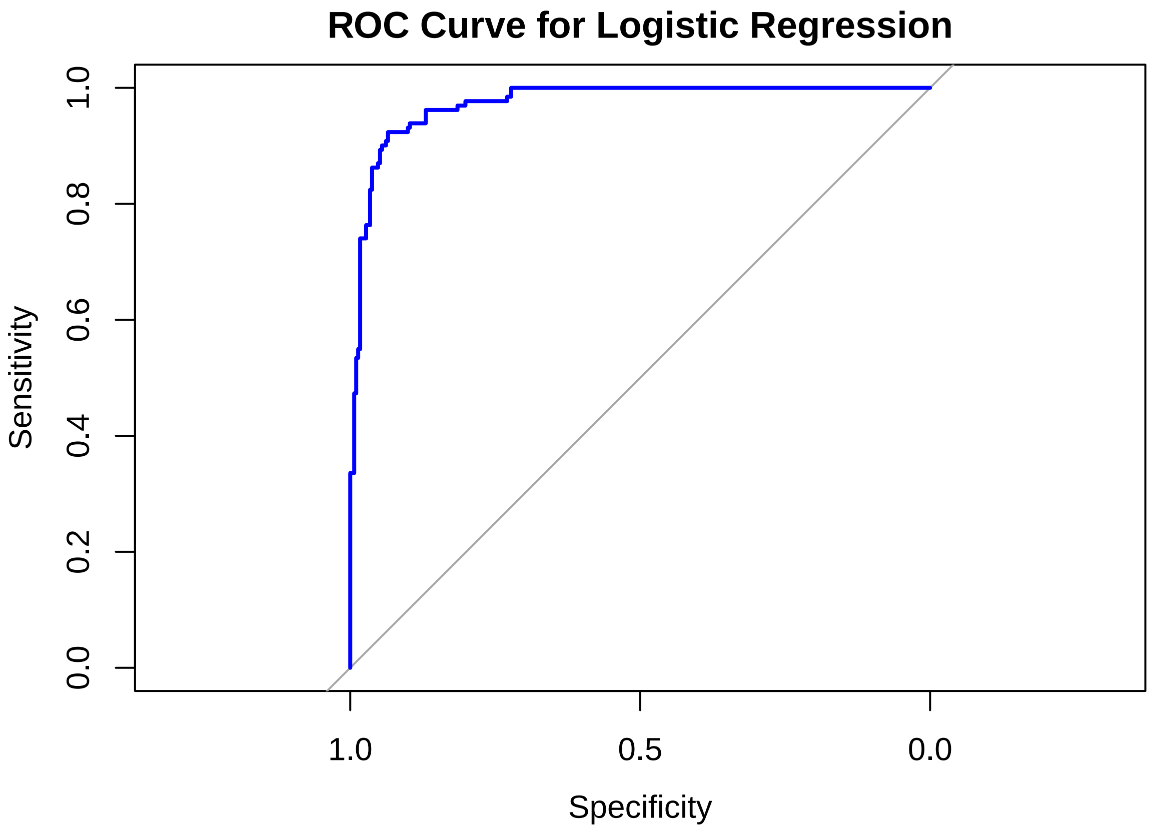
Balanced Accuracy: 0.895

Sensitivity (detecting high life expectancy correctly): 0.966

Specificity (detecting low life expectancy correctly): 0.824

Interpretation: Model performs very well, slightly better at identifying countries with higher life expectancy.

```r
# AUC
roc_obj <- roc(as.numeric(test$low_lifeexp)-1, pred_prob)
plot(roc_obj, col = "blue", main = "ROC Curve for Logistic Regression")
```



```
auc(roc_obj)
```

```
## Area under the curve: 0.9736
```

Key Findings

1. Most important predictors: schooling, hiv_aids, adult_mortality, gdp.
2. Higher schooling reduces risk; higher HIV/AIDS prevalence and adult mortality increase risk.
3. LASSO and stepwise selection (AIC/BIC) largely agree on important variables.
4. The final model shows excellent predictive performance (Accuracy 0.92, AUC 0.97).
5. Some warnings about fitted probabilities = 0/1 indicate potential separation, but overall performance remains robust.

Conclusion

This analysis highlights key predictors of low life expectancy and demonstrates effective modeling using stepwise selection and LASSO. The model is highly accurate and can inform health policy priorities, such as improving education and addressing HIV/AIDS prevalence. Thank you!