Assignment A1: Predictive Statistical Modelling in R

Student Name *(as per record)*	Adekola Oluwatade Job	**Student No**	215383256

	Exceptional	Meets expectations	Issues noted	Improve	Unacceptable
Prepare Exec Report					
Prepare Data					
Discover Relationships					
Create Models					
Evaluate & Improve					
Provide Solution					
Research & Extend					
Brief Comments					Total

Executive summary (half page limit)

“Mortality rate under 5 is equivalent to Poverty” And “Electricity Access is Equivalent to wealth”
The target variable selected (Mortality Rate under 5) was based on an extensive research on poverty prevalence in some countries. The problem statement for this assignment is finding target variables in the world bank indicators that predicts or explains the Wealth and Poverty prevalence of people in a nation. Here, the selected variable (Mortality rate under 5) depicts the poverty variable while some selected predictors like access to electricity, improved sanitation, improved water source, crop production index, literacy rate, rural population are all predictors that would try to explain mortality rate under 5.

In recent research articles for instance, it’s been discovered that malnutrition and lack of access to good infrastructures have been a contributing factor to infant deaths. On the other hand, electricity access, also an extensive research is a big determinant of wealth and poverty. Here, electricity access will be used as a target variable for **wealth** while Mortality rate under 5 will be used as a target variable for **poverty.**

Business Problem
Here, the business problem is trying to determine target factors that causes premature deaths among infants under 5 years of age. Here, we establish some variables (predictors) that depicts factors that influences premature infant deaths. Some predictors for this variable have been selected and based on these we’d discover factors that influences mortality of children especially children under 5. Based on some research articles, malnutrition has been deemed a very good contributing factor to this, and hence the selection of some food predictors like crop index, agricultural land, food index etc. Some environmental and social predictors are going to be examined and thus discover factors that have strong correlations with mortality rate under 5. From the insights generated, we can suggest to world bank the possible counter measures to tackle the problem stated.

Solution to Business Problem
After plotting the correlation charts, it was discovered that rural population is highly correlated with infant mortality. Also, food production index is also moderately correlated with mortality rate under 5.

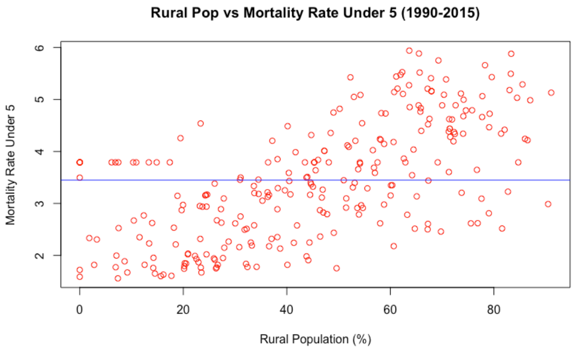
On the other hand, access to improved water and sanitation are highly negatively correlated with mortality rate under 5, which signifies how important good water and sanitation is to a child’s health. But due to poverty of some mothers, there isn’t enough access to these facilities.

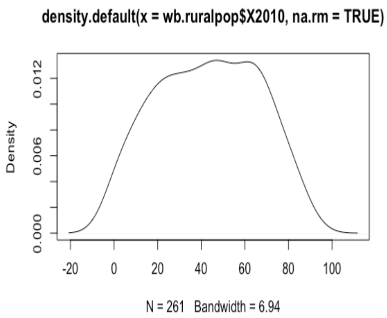
Extension
As a result of the above problems stated, it’s obvious that these high mortality rate under 5 are dominant in rural areas where there’s lack of access to improved water, sanitation, electricity and other social and environmental facilities. Therefore, it’s highly recommended that World bank’s aid should be more supplied to rural areas so as to reduce these problem

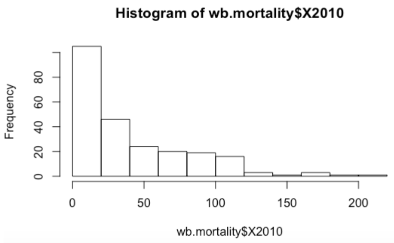
Data exploration and preparation in R (one page limit)

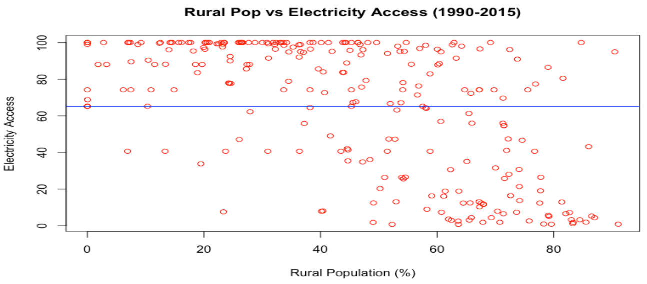
Expectation

Here, the data needed to solve the business problem stated above are data that are actually reflective of poverty e.g. population living in slums and Mortality rate under 5. Firstly, some candidate target variables were selected against some potential predictors for poverty. In the world bank site, the selected candidate target variables are, Electricity Access, Mortality rate of children under 5 and possibly unemployment. Then some of the selected predictors include, population of rural areas, air pollution, net migration, improved sanitation, improved water, access to electricity, health expenditure, central government debt, food production index etc.



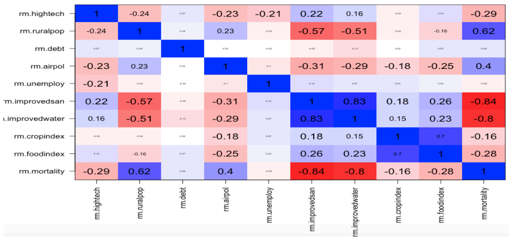


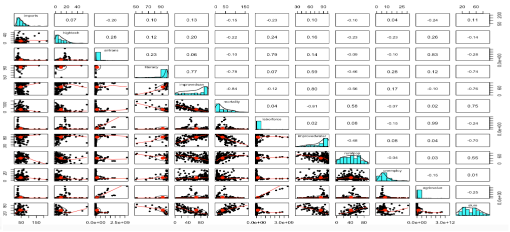




From the scatter plot shown above, it's seen that there is a very positively linear relationship between mortality rate under 5 and rural population. Thus, this gives a basis for explaining poverty in that regard. Also, the correlation plot shows very strong relationships between the predictors and the target variable.

Extension



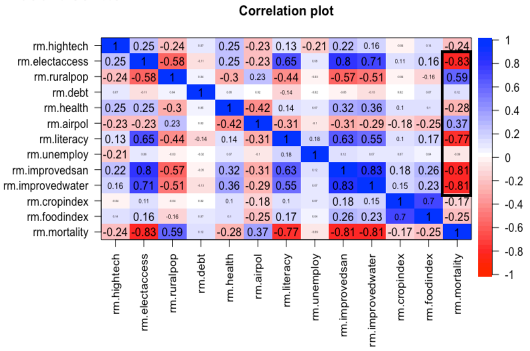


From the above, the target variables, mortality and population in slum were tried side by side but due to the number of NAs in the target variable for slum, mortality rate was chosen instead.

Discovering Relationships and Data Transformation in R (one page limit)

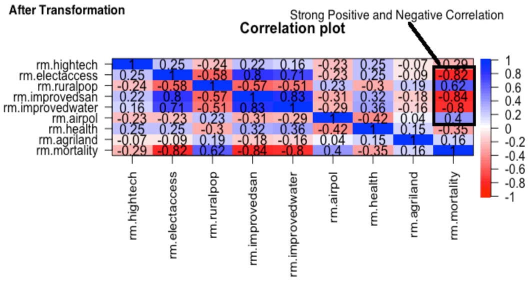
Expectations

Before Transformation



After Transformation

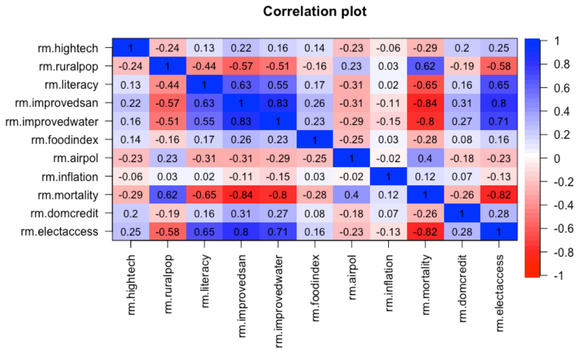
Strong Positive and Negative Correlation



From the above, the target variable was plotted against the selected candidate predictors. Most of the predictors selected were selected based on research articles and the outcome was as a result of the sum of row means of each selected predictor which was used to eliminate the NA values. Also, a Log transformation was done to the target variable because of its unit which improved the correlation plot. The high negative correlation explains so much about how the influence of improved sanitation, improved water and electricity access can have on reduction of mortality rate under 5.

Also, the elimination was due to the fact that not much records were collected for population living in slums and thus allows the usage of mortality rate under 5 as the target variable. This target variable much likely has some high insights to generate through its relationship with other predictors.

Extension

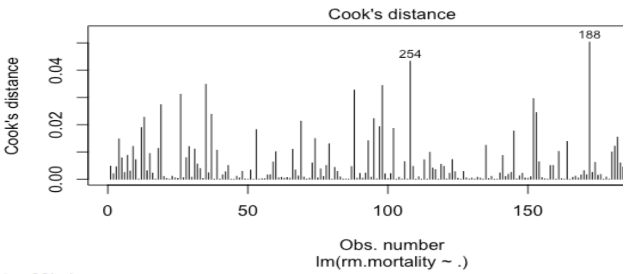


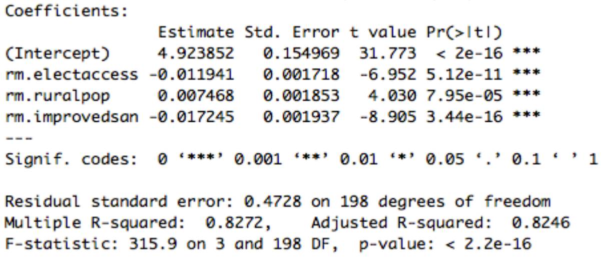
Another very interesting target variable is the electricity access variable. From the correlation plot, it's discovered that there's a strong positive correlation between electricity access and some variables like, literacy rate, improved sanitation, improved water but a strong negative correlation with rural population and mortality rate under 5. We can deduce that this is a good candidate plot to build our model with. Although it's noticed that there is potential multicollinearity relationship between some predictors, this will be ignored for now because some techniques that would be used later on to remove them. Variance inflation factor is a good technique that will be used for removing multicollinearity in this situation. Also, from the percentage of NA values for each predictor, Literacy rate will be removed due to its large number of NAs.

Create Multiple Regression Model(s) in R (one page limit)

Expectation

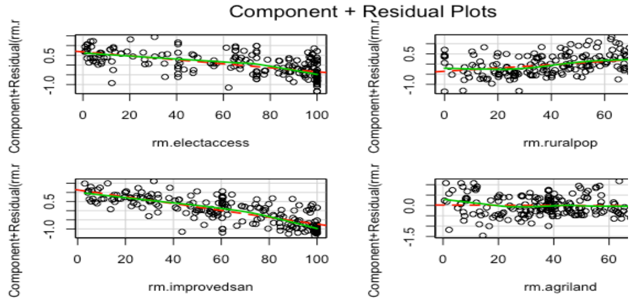
Here, the next step is building a model after discovering relationships between variables and doing the final selection of variables for further analysis. This is initiated with building the final model with row means then putting all selected variables in a data frame. The cook distance test is used in removing extreme values as seen in figure 1 below. After the cook distance test was done, some variables were eliminated based on their P values, VIF and the residual plot. All these can be seen in the figures 1,2,3,4 respectively.

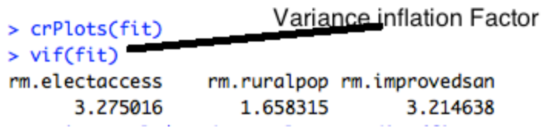




```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.923852   0.154969  31.773  < 2e-16 ***
rm.electaccess -0.011941  0.001718  -6.952 5.12e-11 ***
rm.ruralpop    0.007468   0.001853   4.030 7.95e-05 ***
rm.improvedsan -0.017245  0.001937  -8.905 3.44e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4728 on 198 degrees of freedom
Multiple R-squared:  0.8272,	Adjusted R-squared:  0.8246 
F-statistic: 315.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

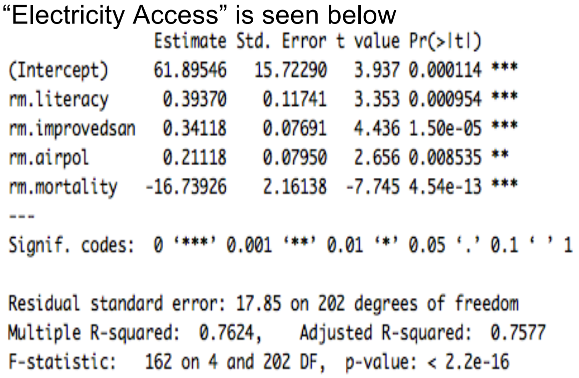




```
> crPlots(fit)
> vif(fit)
rm.electaccess    rm.ruralpop rm.improvedsan 
      3.275016       1.658315       3.214638 
```

Extension

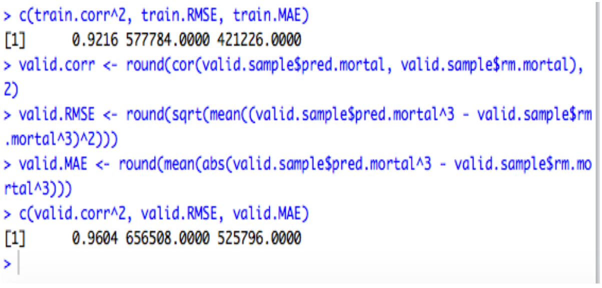
The models built were models for predicting poverty and wealth respectively. The model for "**Mortality rate under 5**" explains poverty while that of "**Electricity access**" explains wealth. Predictive models were built for these target variables. The final predictive model for "Electricity Access" is seen below



```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    61.89546   15.72290   3.937 0.000114 ***
rm.literacy     0.39370    0.11741   3.353 0.000954 ***
rm.improvedsan  0.34118    0.07691   4.436 1.50e-05 ***
rm.airpol       0.21118    0.07950   2.656 0.008535 ** 
rm.mortality  -16.73926    2.16138  -7.745 4.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.85 on 202 degrees of freedom
Multiple R-squared:  0.7624,	Adjusted R-squared:  0.7577 
F-statistic:   162 on 4 and 202 DF,  p-value: < 2.2e-16
```

From the above, the predictive model for the target variable has an R squared value of 0.76 and all P values are less than 0.005. This was done after eliminating some variables based on their high P values. The F statistics is also a good measure of its goodness as it improved its mean by 162. Likewise, 80 percent of the variables were used for training and 20 percent was used for validation. This was then used to predict the outcome of year 2020. The outcome of this prediction is seen below



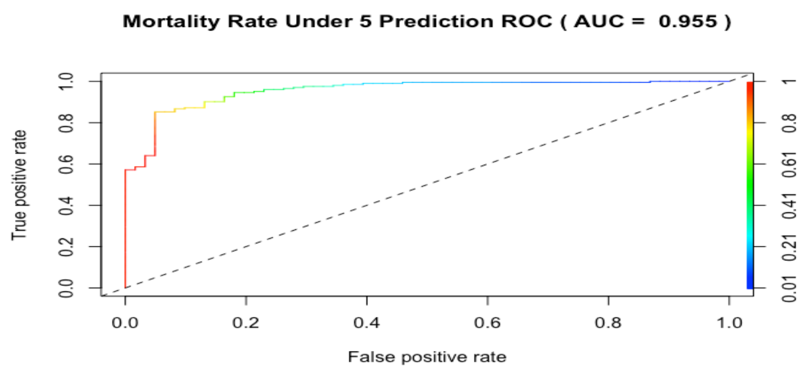
```
> c(train.corr^2, train.RMSE, train.MAE)
[1]      0.9216 577784.0000 421226.0000
> valid.corr <- round(cor(valid.sample$pred.mortal, valid.sample$rm.mortal),
2)
> valid.RMSE <- round(sqrt(mean((valid.sample$pred.mortal^3 - valid.sample$rm
.mortal^3)^2)))
> valid.MAE <- round(mean(abs(valid.sample$pred.mortal^3 - valid.sample$rm.mo
rtal^3)))
> c(valid.corr^2, valid.RMSE, valid.MAE)
[1]      0.9604 656508.0000 525796.0000
> 
```

From the prediction outcome, there's a little variation in the trained data outcome and validation data which explains the goodness of the model and can hence be used to generate further insights.

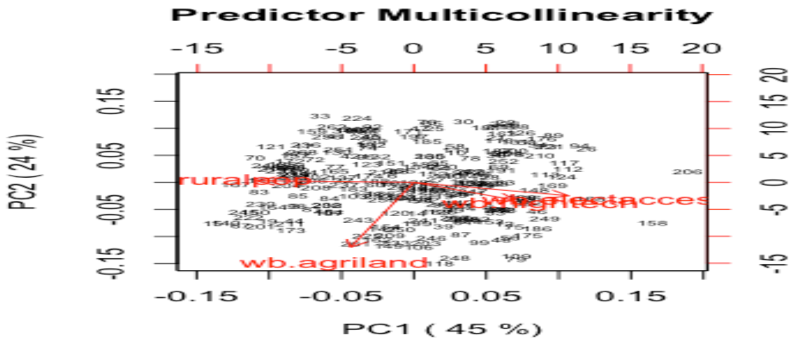
Evaluate and Improve the Model(s) in R (one page limit)

Expectation

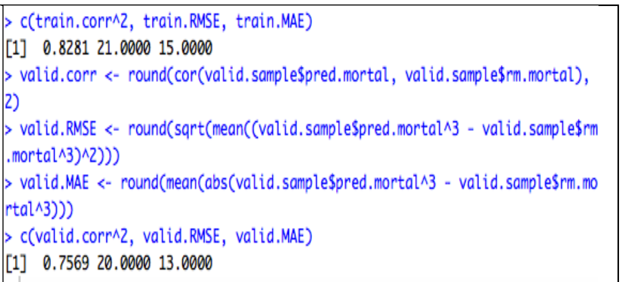
After the model was built, some tests were run on the model to evaluate its performance, check for multicollinearity once more with better graphical means and also using the VIF technique. More variables were eliminated using the LEAPS package that suggests models that do not (fully) pass the Z-test.



The model Performance Plot can be seen from the plot. The plot shows the performance of the built model by plotting the true positive rate against the false positive rate. This is done after the validation have been implemented. From the plot, the true positive rate gives roughly a 0.95 accuracy and a false positive rate of roughly 0.15. This gives the model a very good performance rating. The ROC plot follows the top left axes of the chart which signifies a perfect classifier as seen on the figure shown. Also, it's AUC gives a very large value.

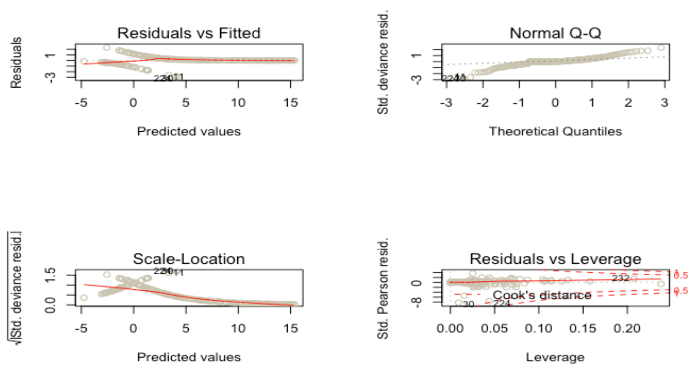


The Big plot command allows a better view of all the multicollinearity between all predictors. As seen in the figure, the distance measure between one predictor to another is quite wide and therefore signalling non-multicollinearity.



```
> c(train.corr^2, train.RMSE, train.MAE)
[1]  0.8281 21.0000 15.0000
> valid.corr <- round(cor(valid.sample$pred.mortal, valid.sample$rm.mortal),
2)
> valid.RMSE <- round(sqrt(mean((valid.sample$pred.mortal^3 - valid.sample$rm
.mortal^3)^2)))
> valid.MAE <- round(mean(abs(valid.sample$pred.mortal^3 - valid.sample$rm.mo
rtal^3)))
> c(valid.corr^2, valid.RMSE, valid.MAE)
[1]  0.7569 20.0000 13.0000
```

The RMSE and MAE is another effective method of measuring model performance. From the figure above, it's seen that that the RMSE value is lower which indicates a better fit for the predictive model and also measures how accurately it predicts mortality rate under 5.



The residual vs fitted plot shows the predicted values as against the model and this shows how the performance of the model is for the seed 2020. All predicted values actually fall into the line of the predictive model and are scattered around the centre line. Also, the scale location has a good measure with the model. For the residual vs leverage plot, after all outliers have been removed using the cook's distance test have also fallen in line with the predictive model.

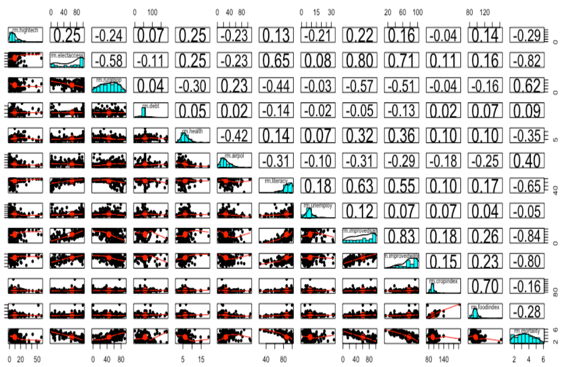
The Normal Q-Q plot (Quantile-Quantile) fall way into the reference line which also suggests its uniform distribution and accuracy.

Provide an Integrated Solution in R (one page limit)

Expectation

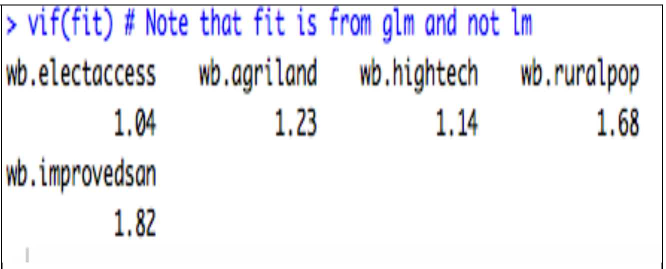
Based on extensive research it is logical to explain why mortality rate is highly correlated with some most predictors selected and also, electricity access being a false target variable for wealth is also highly correlated with most of the independent variables listed. The processes shown below will give an analytic interpretation, a process flow of the analysis done, models built and insights gotten for world bank recommendation.

Figure1



The correlation plot above shows the initial selected predictors against the target variable.

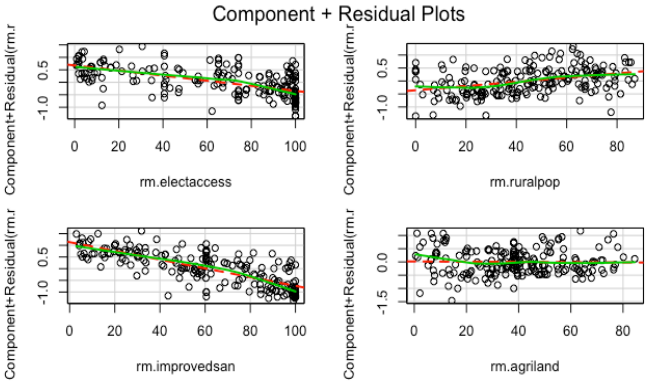
Figure2



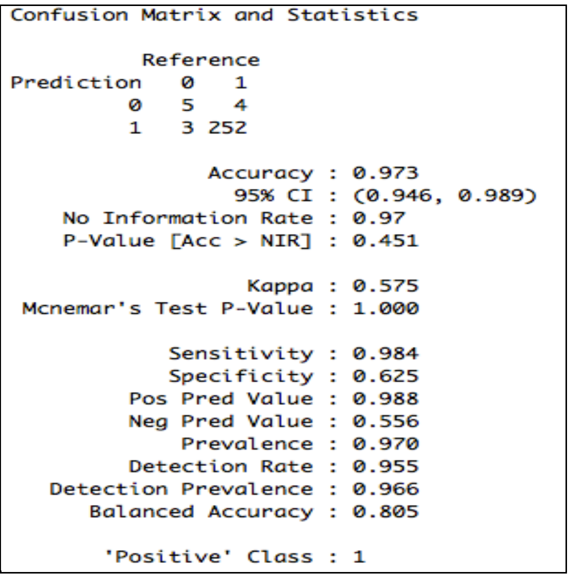
```
> vif(fit) # Note that fit is from glm and not lm
wb.electaccess    wb.agriland    wb.hightech    wb.ruralpop
          1.04           1.23           1.14           1.68
wb.improvedsan
          1.82
```

The VIF summary above shows that there's no multicollinearity relationship between the final selected predictors. This also gives a good basis for the variable selection.

Figure 3



Residual plot for the selected predictors fall the within the model line. This explains its goodness and accuracy. With these 4 predictors, the model was built and model predictive accuracy was tested.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0   5   4
         1   3 252

               Accuracy : 0.973
                 95% CI : (0.946, 0.989)
    No Information Rate : 0.97
    P-Value [Acc > NIR] : 0.451

                  Kappa : 0.575
 Mcnemar's Test P-Value : 1.000

            Sensitivity : 0.984
            Specificity : 0.625
         Pos Pred Value : 0.988
         Neg Pred Value : 0.556
             Prevalence : 0.970
         Detection Rate : 0.955
   Detection Prevalence : 0.966
      Balanced Accuracy : 0.805

       'Positive' Class : 1
```

From Figure 4, the confusion matrix shows how truly a model performs by testing it against previous results. It's seen that the model when tested against known data, it has a 0.973 accuracy level and a kappa value accuracy of 0.97.

Further Research and Extensions in R (one page limit)

Expectation

The logistic regression used in this classification method used allows for easy interpretation of data and good visualisation. As seen in the figure 3 below, the numerical data is converted into categories in order to be classified and thus allow the use of logistic regression. The classification method categorises the mortality rate into two segments i.e. high, medium and low. This gives an insight on how the influence of the selected predictors have on the target variables. As seen from the bar chart below, the red bar signals the percentage of moderate to high mortality rate across the countries represented in the data and the blue.

We'd also use a classification method to classify poverty based on some factors into low, moderate and high. This will provide a better visualisation technique for proper explanation and view to the audience. As seen in figure 1 below, the world map will be a better visualisation method to explain poverty across the world.

In figure 2, it show the number of classes the numerical variable is separated into and also the Kappa value shows the level of agreement between variables used.

Figure 1 (See Code for better view and info)

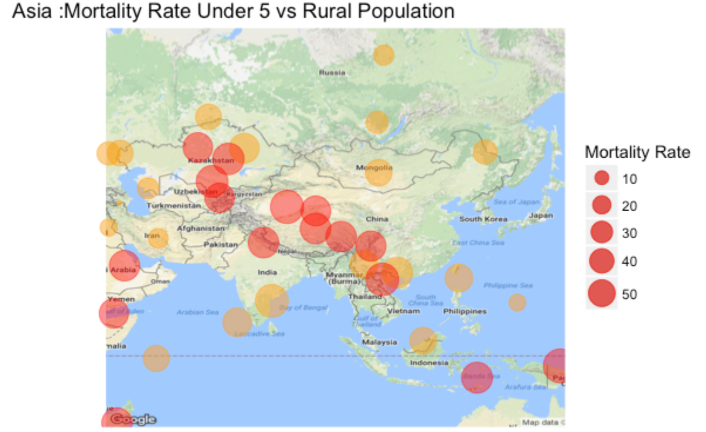
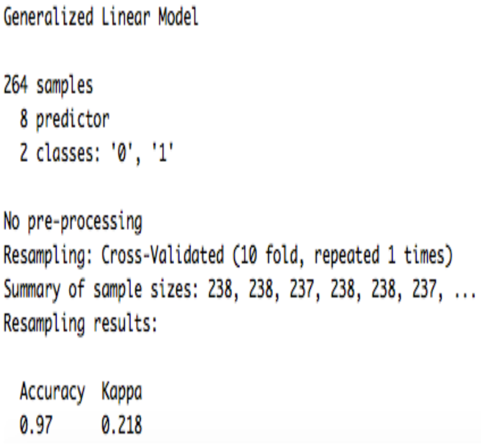
 Asia :Mortality Rate Under 5 vs Rural Population Mortality Rate 10 20 30 40 50 Russia Kazakhstan Mongolia Uzbekistan Kyrgyzstan Turkmenistan Afghanistan Iran Pakistan China South Korea Japan Sea of Japan East China Sea Nepal India Myanmar (Burma) Thailand Vietnam Philippines Philippine Sea South China Sea Gulf of Thailand Bay of Bengal Arabian Sea Laccadive Sea Oman Yemen Arabia malia Malaysia Indonesia Banda Sea Arafura Sea Google Map data ©	The google map shown, shows the distribution of the rate at which mortality rate is impacted according to countries. The bigger the red cycle, the more the mortality rate. This is plotted against rural population in different countries. As seen in the model, rural population is also a predictor for mortality rate. This map could create a basis for decision making and recommendation based on countries with bigger red circles. Based on this, more preventive measures can be taken towards specific nations of the world to combat poverty.

Figure 2



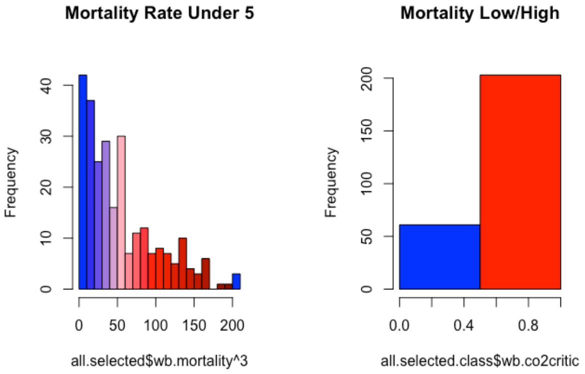
```
Generalized Linear Model

264 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 238, 238, 237, 238, 238, 237, ...
Resampling results:

  Accuracy  Kappa
  0.97      0.218
```

Figure 3



References

1. Jahan, S. (2017). *Poverty and infant mortality in the Eastern Mediterranean region: a meta-analysis*.
2. Washington Post. (2017). *Our infant mortality rate is a national embarrassment*. [online] Available at: https://www.washingtonpost.com/news/wonk/wp/2014/09/29/our-infant-mortality-rate-is-a-national-embarrassment/?utm_term=.ac3724dcf698 [Accessed 9 Apr. 2017].
3. Dornan, M., 2014. Access to electricity in Small Island Developing States of the Pacific: Issues and challenges. *Renewable and Sustainable Energy Reviews*, *31*, pp.726-735.
4. Yceo.yale.edu. (2017). *Electricity and Poverty in Indonesia | Center for Earth Observation*. [online] Available at: http://yceo.yale.edu/electricity-and-poverty-indonesia [Accessed 9 Apr. 2017].
5. Damodar Sahu, A. (2017). *Levels, trends & predictors of infant & child mortality among Scheduled Tribes in rural India*. [online] PubMed Central (PMC). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4510772/ [Accessed 9 Apr. 2017].
6. Moffitt, T., Arseneault, L., Belsky, D., Dickson, N., Hancox, R., Harrington, H., Houts, R., Poulton, R., Roberts, B., Ross, S., Sears, M., Thomson, W. and Caspi, A. (2017). *A gradient of childhood self-control predicts health, wealth, and public safety*.