Data Mining and Analysis (18ECSC301)
Project Report
On
United Nations Millennium Development Goals

Submitted by

Saloni Shah	**01FE18BCS183**
Samarth.M.M	**01FE18BCS184**
Samarth.R	**01FE18BCS185**
Sanjana.K	**01FE18BCS190**

SCHOOL OF COMPUTER SCIENCE & ENGINEERING
HUBLI - 580 031(India)
Academic year 2020-21

INTRODUCTION

The member states of the United Nations agreed to a set of goals to measure the progress of global development. The aim of these goals was to increase standards of living around the world by emphasizing human capital, infrastructure, and human rights.

The eight goals are given below,

1. To eradicate extreme poverty and hunger
2. To achieve universal primary education
3. To promote gender equality and empower women
4. To reduce child mortality
5. To improve maternal health
6. To combat HIV/AIDS, malaria, and other diseases
7. To ensure environmental sustainability
8. To develop a global partnership for development

The UN measures progress towards these goals using indicators such as percent of the population making over one dollar per day.

PROBLEM STATEMENT

The task is to predict the change in the indicators one year and five years into the future. Predicting future progress will help us to understand how we achieve these goals by uncovering complex relations between these goals and other economic indicators. The UN set 2015 as the target for measurable progress. Given the data from 1972 - 2007, we need to predict a specific indicator for each of these goals in 2008 and 2012.
(There is a fair amount of missing data from the training dataset, and strategies need to be devised for dealing with the missing data. Missing values are labeled with NaN)

DATASET

Since its founding in 1944, the World Bank has been gathering data to help it alleviate poverty by focusing on foreign investment, international trade, and capital investment. The World Bank provides these data to the public through their data portal.

The data is available from 1972-2007 on over 1200 macroeconomic indicators in 214 countries around the world. A random snapshot of the data looks like the below. Each row represents a timeseries for a specific indicator and country. The row has an id, a country name, a series code, a series name, and data for the years 1972 - 2007.

Given in the Training Data (TrainingSet.csv)

- Quantitative values of various measures of each country from 1972 to 2007
- Shape of the data = (195402,40)
- Number of Columns = 40
  - 1st column - index
  - 2nd to 37th - years (1972-2007)
  - 38th - Country Name
  - 39th - Series Code
  - 40th - Series Name

A random snapshot of the training data looks like the below.

Reviewing the dataset

	1972 [YR1972]	...	2007 [YR2007]	Series Name	Country Name	Series Code
97510	NaN	...	19	Time to export (days)	Ghana	IC.EXP.DURS
16297	NaN	...	0	Currency composition of PPG debt, Pound sterli...	Azerbaijan	DT.CUR.UKPS.ZS
34357	186000.0000	...	0	PPG, bonds (TDS, current US$)	Botswana	DT.TDS.PBND.CD
126538	NaN	...	54	Newborns protected against tetanus (%)	Jamaica	SH.VAC.TTNS.ZS
30573	NaN	...	NaN	Secondary education, teachers	Bhutan	SE.SEC.TCHR
126818	107.3836	...	NaN	School enrollment, primary, male (% gross)	Jamaica	SE.PRM.ENRR.MA
101060	NaN	...	NaN	Net bilateral aid flows from DAC donors, New Z...	Grenada	DC.DAC.NZLL.CD
18552	NaN	...	NaN	Self-employed, total (% of total employed)	Bahamas, The	SL.EMP.SELF.ZS

Figure 1: Training Data

Given in the file Prediction Data (SubmissionRows.csv)

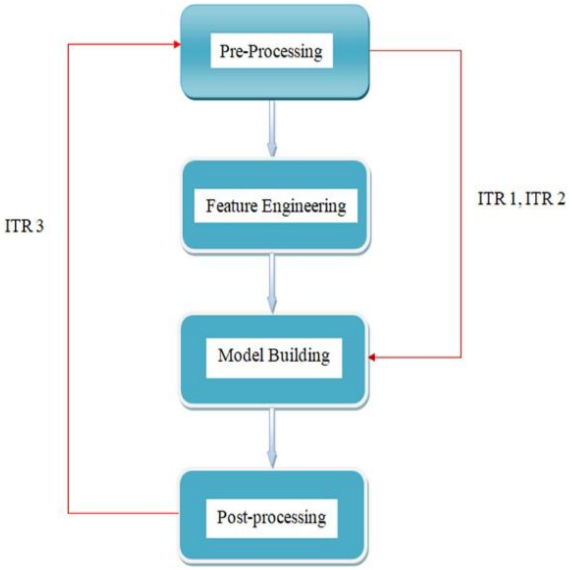
- Index of the columns to be predicted for the year 2008 and 2012
- Shape = (737,2)

A random snapshot of the training data looks like the below.

	2008 [YR2008]	2012 [YR2012]
559	NaN	NaN
618	NaN	NaN
753	NaN	NaN
1030	NaN	NaN
1896	NaN	NaN
1955	NaN	NaN
2090	NaN	NaN
2690	NaN	NaN
3233	NaN	NaN
3292	NaN	NaN

Figure 2: Prediction Data

METHODOLOGY



DATA-PREPROCESSING

We first extracted the data chunks as per the indicators
i.e ensure primary education - 2.1,
reduced child mortality - 4.1,
environmental sustainability - 7.8,
global partnership among nations - 8.16,
eradicating extreme poverty and hunger - 1.2,
combat HIV - 6.1,
combat malaria and other diseases - 6.7,
women empowerment and gender equality - 3.2,
maternal healthcare - 5.1 .

Then for the missing values we went ahead with different techniques to handle them, since ignoring them would have a greater impact on the predicted values.

Last Observation Carried Forward: We first tried to fill with the Last observation carried forward (LOCF) or Forward Filling method. But one caveat about this method was it cannot fill in the missing values present in the beginning of the dataset.

	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
559	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048
1896	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964
3233	0.939	0.939	0.939	0.939	0.939	0.939	0.939	0.939	0.939	0.939	0.939
4570	0.936	0.936	0.936	0.936	0.936	0.936	0.936	0.936	0.936	0.936	0.936
5907	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 3: Imputation using LOCF

Next Observation Carried Backward: Similarly trying with the Next observation carried backward (NOCB) or Backward Filling method does end up in having the missing values in the end. Inorder to handle the missing values at the end, we use the combination of both the methods, i.e Forward and Backward filling method.

	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
1030	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
7715	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
10389	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
11726	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
17074	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Figure 4: Imputation using NOCB

Multivariate Imputation by Chained Equations: We also used the Imputation Using Multivariate Imputation by Chained Equation (MICE) or Iterative Imputation technique to handle the missing values. MICE is a multiple imputation method used to replace missing data values in a data set under certain assumptions about the data missingness mechanism (e.g., the data are missing at random, the data are missing completely at random).

	0	1	2	3	4	5	6	7	8	9	10
0	0.2909	0.072752	0.2272	0.014587	0.242088	0.040157	0.0684	0.094574	0.0194	0.0268	0.123223
1	0.2852	0.070859	0.2219	0.014180	0.239624	0.039117	0.0661	0.091726	0.0184	0.0251	0.120459
2	0.2798	0.068966	0.2151	0.013772	0.236954	0.038061	0.0632	0.088787	0.0173	0.0230	0.117906
3	0.2742	0.067033	0.2062	0.013371	0.234435	0.036989	0.0597	0.087600	0.0163	0.0211	0.115488
4	0.2683	0.065065	0.1962	0.012968	0.232024	0.035899	0.0557	0.084000	0.0153	0.0194	0.113133

Figure 5: Imputation using MICE

These methods contributed while building the learning model with different combinations.

MODEL BUILDING

The models used are-

Linear Regression

Linear regression is one of the most popular models used to predict the values in any field. The model needs a set of data to be trained. The model to be built needs one or many independent

variables and one dependent variable. The model based on the dependent variables, fits a model on the values of all the variables.
Linear regression has a subset called Multiple linear regression. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.

Since we have 36 independent variables which will be used to predict the dependent variable, we will be using the multiple linear regression. The model trains on the training data and optimises the βi values (coefficients).

Using the model trained, the value for 2008 and 2012 is predicted.

Formally, the model for multiple linear regression, given n observations, is

$$y_i = \beta_0 + \beta_{1xi1} + \beta_{2xi2} + \ldots \beta_{pxip} \text{ for } i = 1,2, \ldots n.$$

Model Evaluation

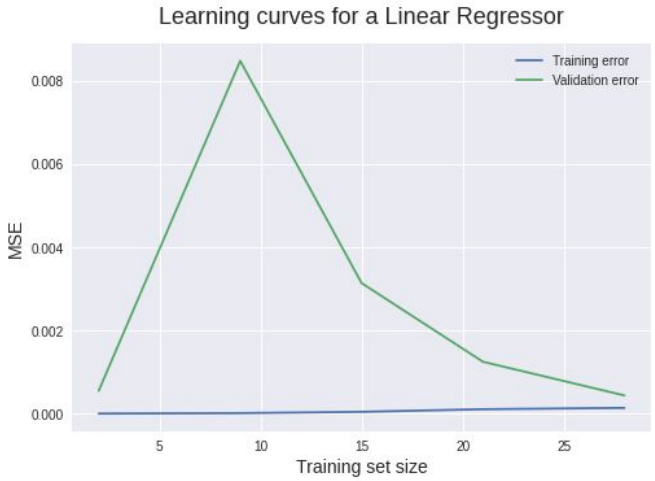


Figure 6: Linear Regression from Forward and Backward Filling method

The Bias Variance curves for Linear regression built with the model preprocessed under Forward and Backward Filling method showed that the mean training scores went upto 0.000137 for the training set and 0.000436 for the validation set as the model complexity increases. We then got an RMSE of 0.1300 for this model.

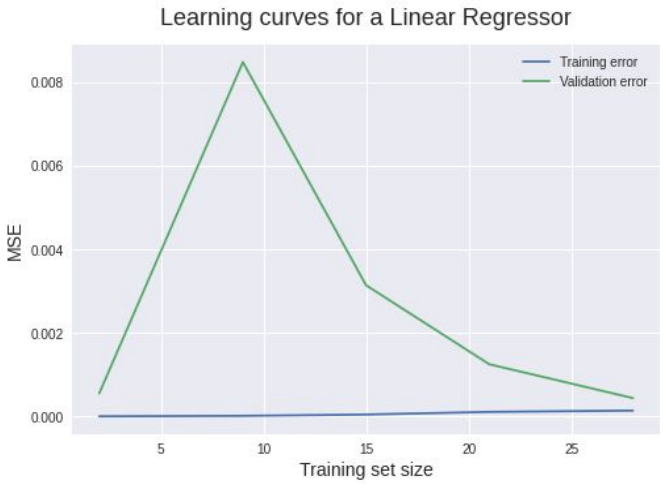


Figure 7: Linear Regression from Iterative Imputation method

Similarly, The Bias Variance curves for Linear regression built with the model preprocessed under Iterative imputation method showed that the mean training scores went upto 0.000137 for the training set and 0.000436 for the validation set as the model complexity increases. We then got an RMSE of 0.0628 for this model.

Ridge Regression

Ridge regression is the variation of the Linear Regression, which helps in reducing the model complexity by preventing overfitting, which may result from simple Linear Regression

$$\textbf{argmin} \sum [y_i - \hat{y}] = \textbf{argmin} \sum [y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_\square x_\square)]^2$$

Here argmin means 'Argument of Minimum' that makes the function attain the minimum. In the context, it finds the β's that minimize the RSS.

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}}{argmin} \|y - XB\|_2^2 + \lambda \|B\|_2^2$$

The term with lambda is often called 'Penalty' since it increases RSS. We iterate certain values onto the lambda and evaluate the model with a measurement such as 'Mean Square Error (MSE)'. So, the lambda value that minimizes MSE should be selected as the final model. This ridge regression model is generally better than the OLS model in prediction. As seen in the formula below, ridge β's change with lambda and becomes the same as OLS β's if lambda is equal to zero (no penalty).

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

We used Ridge regression, with the model having preprocessed with Forward and Backward Filling method by considering all the years, and again considering only the last 3 year for the prediction of the next year. Similarly carried out with the model preprocessed with Iterative imputation.

Model Evaluation

Figure 8: Ridge Regression from Forward and Backward Filling method

The Bias Variance curves for Ridge regression built with the model preprocessed under Forward and Backward Filling method showed that the mean training scores went upto 0.000140 for the training set and 0.000437 for the validation set as the model complexity increases. We then got an RMSE of 0.1300 for this model.

Figure 9: Ridge Regression from Iterative Imputation method

The Bias Variance curves for Ridge regression built with the model preprocessed under Iterative imputation method showed that the mean training scores went upto 0.000140 for the training set and 0.000437 for the validation set as the model complexity increases. We then got an RMSE of 0.0545 for this model.

ARIMA Model (AutoRegressive Integrated Moving Average)

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.
Since, the data provided is non-Stationary, we use ARIMA model instead of ARMA model.
The term 'Auto Regressive'(AR) in ARIMA means it is a linear regression model that uses its own lags as predictors.'p' is the order of AR term. It refers to the number of lags of Y to be used as predictors.
The value of d is the minimum number of differencing (subtract the previous value from the current value) needed to make the series stationary. And if the time series is already stationary, then d = 0.
The term 'Moving Average'(MA) in ARIMA refers to the number of lagged forecast errors that should go into the ARIMA Model. 'q' is the order of MA term.

In our model, we used two different combinations of ARIMA and pre-processing techniques.

1) Global Average
   Averages is considered Globally (all the countries without segregation) with linear interpolation for pre-processing.

2) Continental Average
   Averages is considered Continent (country wise segregation) wise with linear interpolation for pre-processing.

Model Evaluation

The data for this model was preprocessed once using the interpolation techniques, and the RMSE values recorded for each of the variation are as follows:
Global Average: 0.0505
Continental Average: 0.0615

The other preprocessing technique used to train on ARIMA model was backward and forward filling and the root mean squared error are as follows:
Global Average: 0.0493
Continental Average: 0.0494

Since the main distinguishing factor was supposed to be indicators we have plotted the values against the years for each indicator considering some countries.

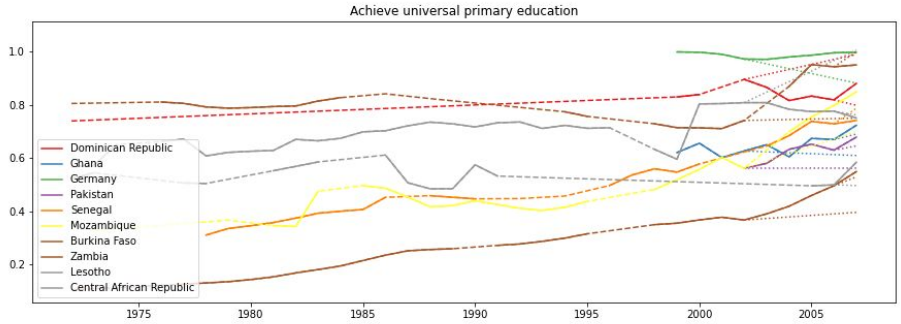


Figure 10: Values against the year for the indicator- To achieve universal primary education

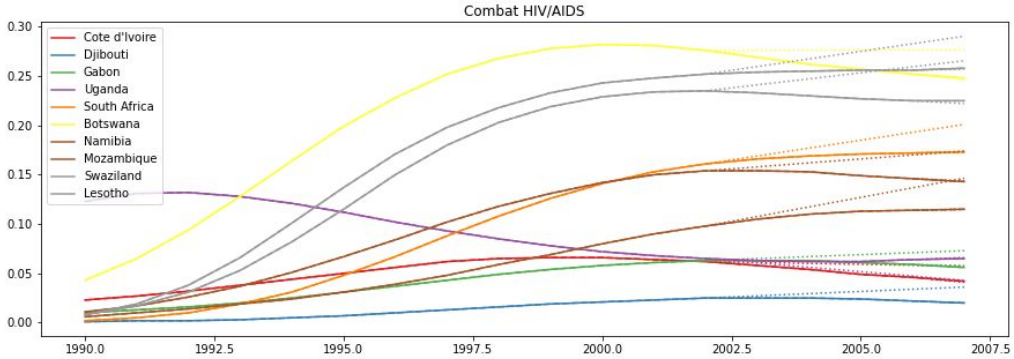


Figure 11: Values against the year for the indicator-To combat HIV/AIDS, malaria, and other diseases

Vector AutoRegression

Vector autoregression (VAR) is a statistical model used to capture the relationship between multiple quantities as they change over time. VAR is a type of stochastic process model. VAR models generalize the single-variable (univariate) autoregressive model by allowing for

multivariate time series.

Like the autoregressive model, each variable has an equation modelling its evolution over time. This equation includes the variable's lagged (past) values, the lagged values of the other variables in the model, and an error term. VAR models do not require as much knowledge about the forces influencing a variable as do structural models with simultaneous equations. The only prior knowledge required is a list of variables which can be hypothesized to affect each other over time.

Model Evaluation

Similar to ARIMA, we tried VAR by preprocessing the data in two different ways. One using linear interpolation and the other using backward and forward filling.

The RMSE values fall as below:

Using linear interpolation: 0.0493

Using backward and forward filling: 0.0494

The plots obtained for the VAR model for some countries showing the change in values with respect to the indicators against the years are added below.

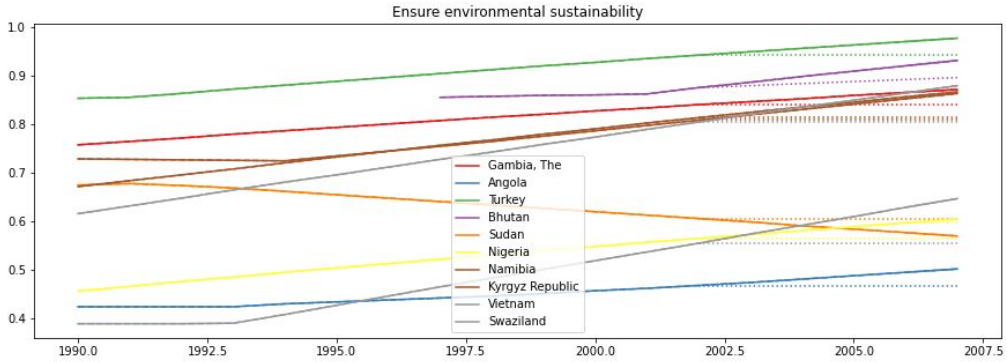


Figure 12: Values against the year for the indicator- To ensure environmental sustainability

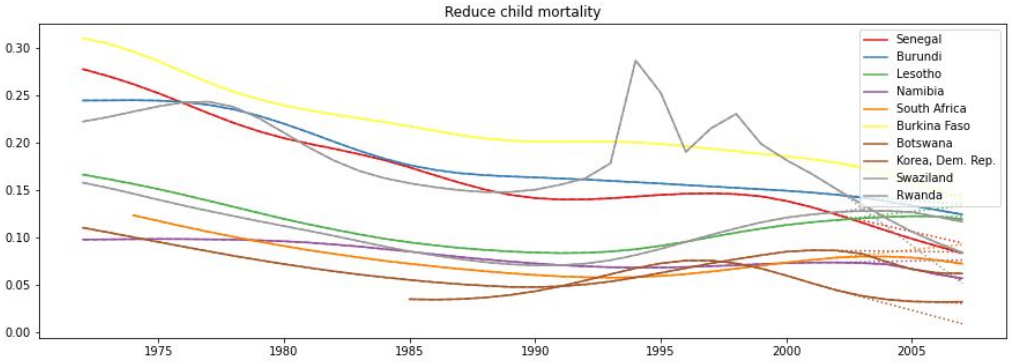


Figure 13: Values against the year for the indicator- To reduce child mortality

Modelling Using Alpha Beta-Filter

An alpha beta filter presumes that a system is adequately approximated by a model having two internal states, where the first state is obtained by integrating the value of the second state over time. Measured system output values correspond to observations of the first model state, plus disturbances.

Mechanical systems where position is obtained as the time integral of velocity. Based on a mechanical system analogy, the two states can be called *position x* and *velocity v*. Assuming that velocity remains approximately constant over the small time interval ΔT between measurements, the position state is projected forward to predict its value at the next sampling time using equation 1.

$$(1) \quad \hat{\mathbf{x}}_k \leftarrow \hat{\mathbf{x}}_{k\ 1} + \Delta \mathrm{T}\ \hat{\mathbf{v}}_{k\ 1}$$

Since velocity variable *v* is presumed constant, its projected value at the next sampling time equals the current value.

$$(2) \quad \hat{\mathbf{v}}_k \leftarrow \hat{\mathbf{v}}_{k-1}$$

If additional information is known about how a driving function will change the *v* state during each time interval, equation 2 can be modified to include it.

The output measurement is expected to deviate from the prediction because of noise and dynamic effects not included in the simplified dynamic model. This prediction error *r* is also called the *residual* or *innovation*, based on statistical or Kalman filtering interpretations

$$(3) \quad \hat{\mathbf{r}}_k \leftarrow \mathbf{x}_k - \hat{\mathbf{x}}_k$$

Suppose that residual *r* is positive. This could result because the previous *x* estimate was low, the previous *v* was low, or some combination of the two. The alpha beta filter takes selected *alpha* and *beta* constants (from which the filter gets its name), uses *alpha* times the deviation *r* to correct the position estimate, and uses *beta* times the deviation *r* to correct the velocity estimate. An extra ΔT factor conventionally serves to normalize magnitudes of the multipliers.

$$(4) \quad \hat{\mathbf{x}}_k \leftarrow \hat{\mathbf{x}}_k + (\alpha)\, \hat{\mathbf{r}}_k$$
$$(5) \quad \hat{\mathbf{v}}_k \leftarrow \hat{\mathbf{v}}_k + (\beta/[\Delta \mathrm{T}])\, \hat{\mathbf{r}}_k$$

The alpha and beta coefficients refer to the two scaling factors, where g is the scaling we used for the measurement, and h is the scaling for the change in measurement over time.

Model Evaluation:

The model was first trained on each country separately which predicted the new values extremely well, with a Root-Mean Squared Error of 0.0484.

Later the model was trained on all the countries together which predicted the values for 2008 ad 2012 with RMSE of 0.0480.

Plots obtained from the values of some indicators are added below.

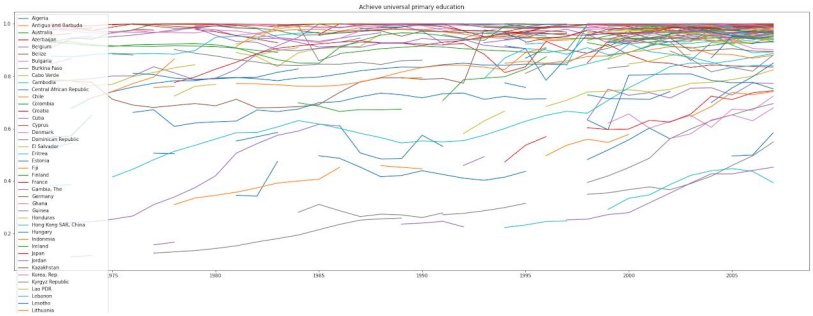


Figure 14: Values against the year for the indicator- To achieve universal primary education

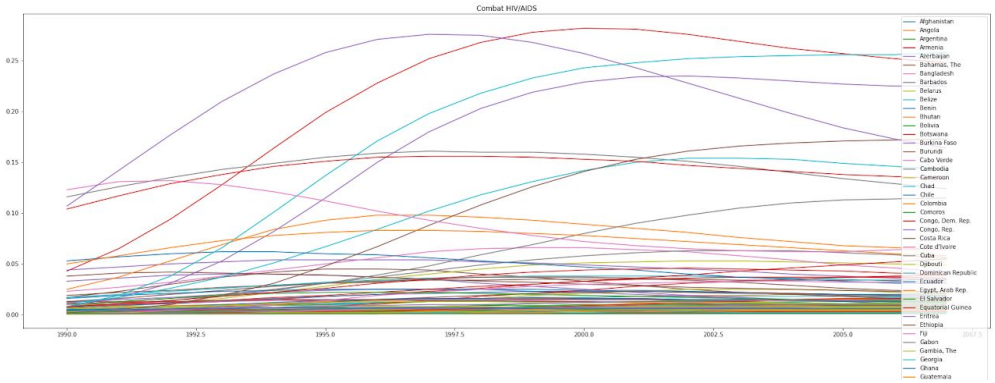


Figure 15:Values against the year for the indicator-To combat HIV/AIDS, malaria, and other diseases

RESULTS

Linear Models like Linear Regression and Ridge Regression did considerably well on the data. ARIMA model being a model preferred for time series data, did much better to predict the values for the years 2008 and 2012. A further variation of ARIMA used, Vector AutoRegression provided a lesser RMSE score with different preprocessing techniques.
Using the final model which uses a variation of Kalman-filter which is alpha-beta filter or also known as gh filter, we stand 18th (as of 2nd January, 2021), with RMSE score of 0.0480.

CONCLUSIONS

Since there were many attributes with missing values, the crucial step was to fill in the missing data, for which we have used different techniques. Indicators being the factor determining the progress, we considered it to be one of the important attributes.

We experimented with different preprocessing techniques over different models used, providing combination of solutions.

REFERENCES:

1. https://www.drivendata.org/competitions/

2. https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779