Health & Economy Analytics on World Bank

Aarya Vardhan Reddy Paakaala[1] Chandrakanth Reddy Nomula[2]
[1,2]MVSR Engineering College, India

Abstract— World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially-recognized international sources. It presents the most current and accurate global development data available and includes national, regional and global estimates. This statistical reference includes over 800 indicators covering more than 150 economies. This paper aims at preprocessing and grouping the WDI of India and USA into health and economy datasets respectively and conducting data analytics on them. Multiple linear regression is performed on the health-based and economy-based indicators of India and USA respectively, to generate predictions and compare them with actual values to understand the concept behind the multiple linear regression model. Additionally, data visualization tool is used to provide a graphical understanding of the data over the years. These results can be useful for businesses, NGOs to find new opportunities and also helps normal people to better understand various data indicators for India and USA.

Key words: Multiple Linear Regression, WDI

I. INTRODUCTION

To understand a country development, we require information about its past and current aspects. People rely on news, internet for such information. The sum of facts printed in newspapers over the year will be very much less than the actual size of facts. Therefore most relevant data to understand world development is world development indicators. The annual publication is released in April of each year. The online database is updated three times a year.

Searching all the data using excel sheets is inefficient and it becomes complex when we want to compare data from various indicators of distinct countries. Storing this huge data in a database is inefficient and performing queries gives data again in the form of huge tables. Effective visualization of data helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Moreover, it is observed that machine learning is applied extensively in our day-to-day life and this trend continues to expand which rises a need for people to understand the techniques of machine learning. Thus, application of multiple linear regression on WDI to predict desired indicator values and compare them with actual values is done to understand how multiple linear regression, a machine learning algorithm handles the data.

This paper mainly focuses on WDI of United States of America and India only. All the results are encapsulated in a shiny app to hide the complex computations from users.

II. METHODOLOGY

A. Data Preprocessing

The first phase is data preprocessing. The WDI dataset is preprocessed by removing null values, selecting required indicators, selecting the years for which data needs to be extracted, selecting the specified countries. Accordingly, the data is grouped into 4 datasets i.e Health-based indicators of USA, Health-based indicators of India, Economy-based indicators of USA, Economy-based indicators of India.

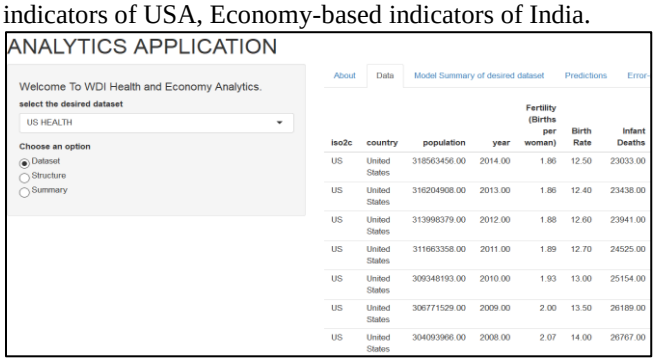

Fig. 1: Preprocessed Health-based Indicators Dataset of USA

B. Multiple Linear Regression

Linear regression is a method which is used to predict the outcome of a variable, the output, or dependent variable, by using a set of independent variables. As the name implies, the method of prediction is linear, with a linear relation of the independent variables being used to predict the value of the outcome variable [1]. In statistics, linear regression models the relationship between a dependent variable and one or more explanatory variables using a linear function. If two or more explanatory variables have a linear relationship with the dependent variable, the regression is called a multiple linear regression. The basic mathematical model of multiple linear regression analysis. [2]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$$

Training the model is the next phase in which the datasets are split into training and testing sets. Once the split is complete, the training set is trained rigorously by multiple linear regression model and a trained model is built which serves as a base for next phase. Once the training model phase is complete, the next phase is predicting phase where predictions are done on the testing data using the trained model. The predicted values are compared with actual values and the error rates are computed using Root mean square error (RMSE).

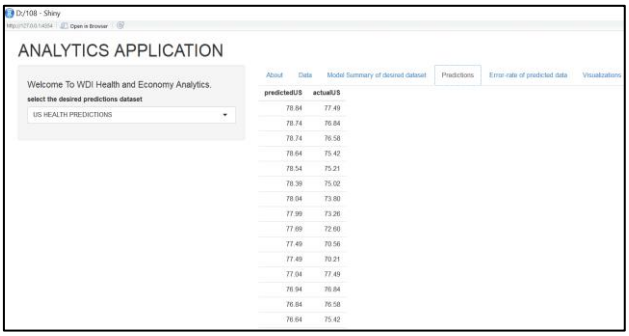

Fig. 2: Predicted & Actual values of Health-based indicators of USA

C. GGPLOT2

For visualization tools to support exploratory data analysis, they must allow analysts to rapidly create and refine visualizations. Some current systems let developers build custom interactive visualizations for the web, but the time and effort required to use these tools can break the flow of an ad hoc analysis session. In contrast, tools such as Tableau (www.tableau.com) and ggplot2 (a visualization package for the R statistical computing language; ggplot2.org) support iterative and interactive construction of graphics using a high-level grammar for concise specification.[3]

ggplot() function consists of data, aesthetics and geometric parameters. Data represent the dataset on which visualizations are to be made. Aesthetics represent the indicators that are to be represented on the x and y-axes. Geometrics represent the geometry of the visualization to be produced.

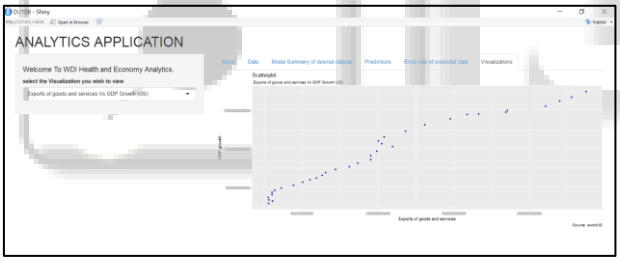

Fig. 3: Visualization Displaying a Relationship between Exports & GDP Growth of USA over the Years

D. Shiny R

Shiny is an R package that makes it easy to build interactive web apps straight from R Studio. Structure of a Shiny app consists of 3 parts. They are a user interface object, a server function and a call to the Shiny app function. The shiny app consists of multiple tabs which are used to interact with the user, compute and display the user's desired output.

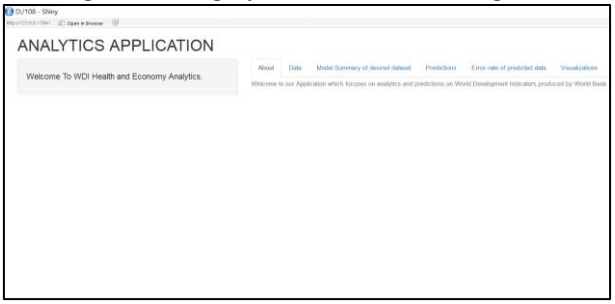

Fig. 4: Overview of Shiny application

E. Requirements

- CRAN R
- R studio
- World Development Indicators dataset

III. PROPOSED SYSTEM

The designed system enables the user perform the following functionalities:

- Perform multiple linear regression on World Development Indicators (WDI) to analyze its mechanism by predicting health-based and economy-based indicators of India and USA respectively.
- —Visualizations of Health-based and Economy-based indicators of USA and India respectively.
- Building a Shiny app and encapsulating all the back-end computations in it.
- More functionalities are proposed to be added to this system such as:
- Providing advanced data visualizations for better and accurate understanding of the data.
- Applying other machine learning algorithms like random forests etc. could be implemented to understand their mechanisms.

IV. CONCLUSION

Implementation of Data visualization plots such as scatterplot and correlation plot is an ideal setting for effective analysis tool to a wide range of experience and non-experienced people. Shiny package helped to further minimize and better organize the application. Data predictions from multiple linear regression are used to predict how the indicators vary over the course of time which would help us interpret the multiple linear regression.

REFERENCES

[1] https://www.codeproject.com/Articles/1172253/The-Price-of-Wine-as-predicted-using-Linear-Regres
[2] A forecast for bicycle rental demand based on random forests and multiple linear regression Found in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) By YouLi Feng, ShanShan Wang, Issue Date: 2017-05
[3] A High-level Language for Interactive Data Visualization, by Leila De Floriani, Issue Date: 2017-04

