Artificial Intelligence: Machine Learning

Lars Askegaard, Bennet Botz, Kaz Matsuo
Department of Mathematics, Statistics, and Computer Sciences, St. Olaf College
Askega2@stolaf.edu, Botz2@stolaf.edu, Matsuo1@stolaf.edu

Abstract

Linear regression is a machine learning model whose purpose is to predict the result of a continuous variable based on one or more input features. In this research paper, we propose two linear regression models. Our first linear regression model predicts the human development index (HDI) based on four input features: extreme poverty, handwashing facilities, percent of population fully vaccinated, and gross domestic product (GDP) per capita. Our second linear regression model predicts the stringency index based on four input features: total cases, total vaccinations, total deaths, and total boosters. To measure the accuracy of our machine learning regression model we used the metrics mean squared error, root mean squared error, and R-squared score. Our results indicate that machine learning can be an accurate tool when making predictions for HDI and stringency index

Introduction

Covid-19 has had a significant impact throughout the entire world. There are a variety of features that impact a countries ability to respond effectively to the virus. Johns Hopkins University Center for Systems Science and Engineering has been managing a global Covid-19 data set. The data is compiled from a variety of sources including the World Health Organization, European Centre for Disease Prevention and Control, BNO News, and local health agencies. There are 238,000 data entries recording relevant information over 68 categories including cases, deaths, vaccinations, and other relevant Covid-19 information. In this study we focused on human development index (HDI) and the stringency index.

HDI is used to rank countries based on their human development. It is a composite index that measures three basic dimensions of human development. These factors are a long and healthy life, knowledge, and a decent standard of living.

The stringency index is a measure of the strictness of public health and social measures implemented by a government in response to the COVID-19 pandemic. It is intended to provide a standardized way to compare the severity of the measures taken by different countries or regions in response to the pandemic. The stringency index is based on a scale from 0 to 100, with higher values indicating more stringent measures. It is calculated using a composite measure that considers nine different types of measures, including school and workplace closures, restrictions on public gatherings, and travel ban

In hopes to gain a deeper understanding of HDI and stringency index we created two machine learning models using linear regression. This study predicts HDI based on extreme poverty, handwashing facilities, percent of population fully vaccinated, and GDP per capita. This study also predicts stringency index based on total cases, total vaccinations, total deaths, and total boosters.

The goal of the study is to develop a linear regression model that predicts the Human Development Index based on the four factors stated above as well as develop several stringency index models based on a low, mid, or high HDI. Using covid metrics such as total cases, total deaths, and total vaccines, we can then predict the countries response to covid given their HDI levels. Different HDI leveled countries could possibly act in different ways and the regression model should provide insight based on past responses of various countries. In order to get a better correlation between the variables several countries represent each category in the high, medium, and low HDI groups. The High group is represented by models of Italy, USA, and Canada, the medium group represented by Brazil, and Colombia, and the Low group represented by Cambodia and Pakistan.

Background: Machine Learning with Linear Regression Techniques

Machine learning is a branch of artificial intelligence and computer science which focuses on how data and algorithms can imitate the way that humans learn and interpret data, and gradually improve its accuracy of understanding through iteration (IBM Cloud Learn Hub). Machine learning is a process in which a computer builds a model based on observed data and uses the model it has built as a hypothesis

to predict pieces of information about the world (Russell and Norvig, pg. 1201). The machine uses this generated model, in combination with visualization software, to numerically solve and discover correlations in real-world problems. Through using various statistical methods of classification and prediction, algorithms uncover pivotal insights in higher and lower volume data mining projects.

When the desired output of our machine learning model is one element among a finite categorical group of values, the ML process is called classification (Russell and Norvig, pg. 1204). A common application of the classification technique is text classification, in which machine learning models strive to correlate text-based descriptions with categorical groups. When researchers desire to predict an original numerical value that does not fit into a categorical framework, we consider this type of learning as regression (Russell and Norvig, pg. 1204). Developing numerical predictive trends, interpreting past numerical data, or relating interdependencies of numerical data points are all common applications of machine learning regression models.

In this publication, we study the two different effects of how the covid metrics of percent of population fully vaccinated, amount of hand washing facilities, extreme poverty, and GDP per capita affect stringency index and human development index. Since stringency index and human development index (HDI) are both continuous numerical scales, we will build regression-based machine learning models to predict numerical trends that exist between these variables.

For this study, we elected to narrow our investigation to relationships which exhibit linear correlation between the observed and measured variables. We utilize the 'LinearRegression' function from the 'Linear_model' package of Scikit-Learn's open-sourced python library to form two different models that predict human development index and stringency index values as linear functions of their independent variables. LinearRegression is an ordinary least squares algorithm for linear regression that fits a linear model based on minimization of the sum of distances between dataset observations and values predicted by linear approximation (*Ordinary Least Squares Linear Regression*). In simpler language, this algorithm strives to minimize the vertical distance between a line of best fit and all data that the algorithm is given access to as training data points. By minimizing the vertical distance between each data point and the predicted regression line, we produce a model that accurately captures correlations in linear relationships. We test the accuracy of this model by randomly separating the data into a 'test' and 'train' datasets which enable us to determine the relative accuracy of the model's prediction for the entire dataset.

Before attempting to fit a linear model to the dataset, it is crucial to determine whether there is a compelling correlation between the independent and dependent variables. In the data preprocessing stage, we illustrate the steps prior to fitting a linear regression model that are necessary for determining the correlative strength of each variable of interest. Once data is properly preprocessed, we apply a training algorithm on a portion of our original data (X_Train) and test with the rest of the original data (X_test) and see how accurate the model is by comparing with the original y and these fitted values (y_pred).

Data Preprocessing Techniques

Preprocessing is an important step when developing a machine learning model. Preprocessing involves cleaning the data before inputting it into the machine learning model. Preprocessing ensures that the data is in the correct format for the machine learning model and increases the accuracy of the model. When preprocessing the covid data for the machine learning model, we first worked on feature engineering, then we removed NAs, and then we normalized data.

Feature engineering involves creating new columns out of existing columns in the data set. We wanted a column that represented the percent of the population that is fully vaccinated. We created a new column titled 'percent_fully_vaccinated'. Our new column 'percent_fully_vaccinated' was created by dividing the columns 'people_fully_vaccinated' and 'population'. We later use the variable 'percent_fully_vaccinated' to predict the HDI.

Next, we removed all the NAs from the dataset. To do this we used the '.notna()' function. This function can be used to remove missing values from a Pandas data frame. If any of our columns of interested contained a NA value, we removed that row from the data set.

In the normalization phase, a few modifications were made in order to account for the presence of outliers and missing data entries. The scaling process which is the stage where the data is scaled to have a mean of 1 and a standard deviation of zero for a better performing model. A standard scaling process, however, is sensitive to outliers and extreme values which most of our data has. To avoid this the preprocessing used Robust scaling where the data is scaled to a median of 0 and an interquartile range of 1. Because the standard scaling scales to a mean the outliers play a strong presence but because the robust method scales to a median, the extreme values minimally affect the scaled dataset.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Equation 1. Standard Scalar

$$X_{new} = \frac{X - X_{median}}{IQR}$$

Equation 2. Robust Scalar

Predicting Human Development Index

Our first Linear Regression model predicts the HDI based on four factors: extreme poverty, handwashing facilities, percent of population fully vaccinated, and GDP per capita. Our goal for this model was to use multiple linear regression to examine how HDI depends on the four factors above. After performing the data preprocessing described in the section above, we began to work on training a model to the data. We split the data into a training set and a test set using the 'test_train_split' function. The training data set was used to train the linear regression models, and the test data set was later used to evaluate the performance of the models. We then trained and tested the models. To evaluate the performance of the data we used the mean square error as well as R-squared. The mean square error measures the difference in the predicted values the actual values in a data set. The R-squared is the variance in the data explained by our model. We achieved a mean squared error of 0.05 and an R-squared value of 0.85. Figure 1 below is a visualization of the actual data points and the data points predicted by our model. The actual data points are depicted in blue and the predicted data points are depitcted in orange.

We noticed that 'percent_fully_vaccinated', 'handwashing_facilities', and 'gdp_per_capita' are positively correlated to HDI and 'extreme_poverty' is negatively correlated to HDI. As the percent of people fully vaccinated increases the HDI also increases. This suggests a possibility that countries with more people fully vaccinated have a higher HDI.

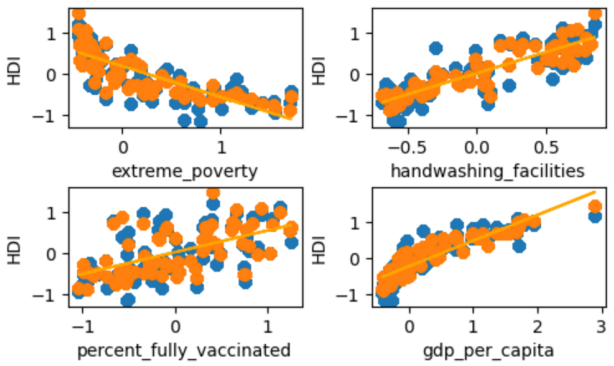

Figure 1. Human Development Index vs. various independent variables after normalization

Predicting Stringency Index

Our second regression model is another Linear Regression that predicts the stringency index based on four factors: Total Cases, Total Deaths, Total Vaccinations, and Total Boosters. Our goal with this model is to see how the various implications of covid changed the stringency, or how restrictive the countries in terms of school closings, shifting to remote work, cancelation of in-person events, and quarantining. Based on our previous model of Human Development Index, we wanted to compare countries in three categories of Human Development Index: High, Middle, and Low. Something to note as a limitation to this method is that low HDI countries generally do not have ample amounts of data that could be used in the regression model as many data entries are missing. Regardless, the model should indicate the correlation of stringency index based on multiple factors of Covid-19.

It seems that the general trend of the High HDI countries, USA, Canada, Italy has a gradual stringency index decrease to the increase in total deaths and total cases. These countries seemed to have a plan that allowed their country to gradually loosen restrictions while the mid/low HDI groups seems to have a sharp drop after a certain number of total cases and deaths. In terms of the vaccination and boosters each country had a similar relation to the stringency index with no significant difference between the given test groups. Regardless of the cross-group analysis, the model itself performed consistently well where the lowest R-squared value was 0.892 and highest at 0.971 and MSE being between 0.012 to 0.222. The slightly high MSE values for Brazil and Pakistan may be a result of overfitted values and could lead to poor predictions as a model.

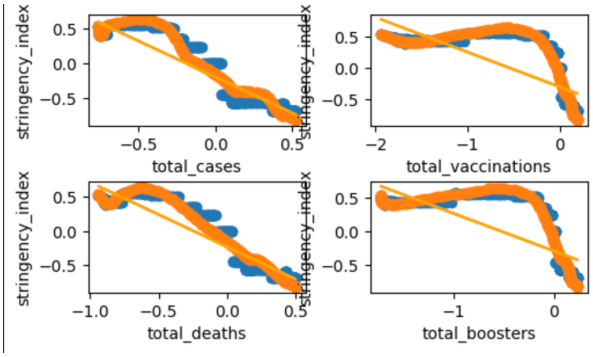

Figure 2. Italy (High) Stringency Index vs. various independent variables after normalization. R^2 = 0.919, MSE = 0.0188, RMSE = 0.137

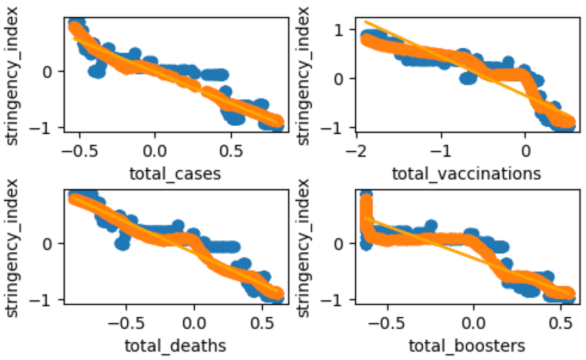

Figure 3. USA (High) Stringency Index vs. various independent variables after normalization. R^2 = 0.953, MSE = 0.012, RMSE = 0.109

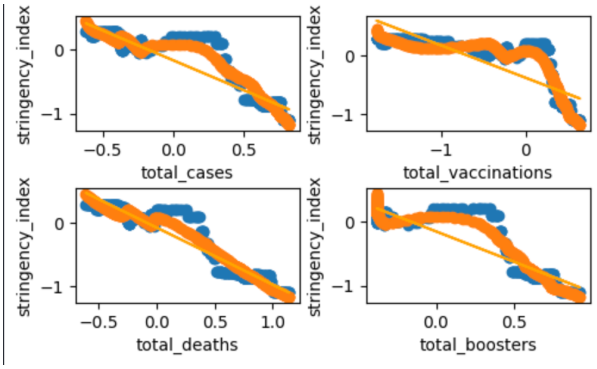

Figure 4. Canada (High) Stringency Index vs. various independent variables after normalization. R^2 = 0.919, MSE = 0.0188, RMSE = 0.137

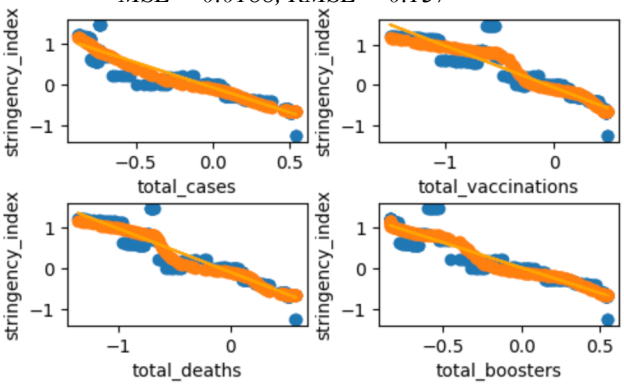

Figure 5. Brazil (Mid) Stringency Index vs. various independent variables after normalization. R^2 = 0.892, MSE = 0.050, RMSE = 0.224

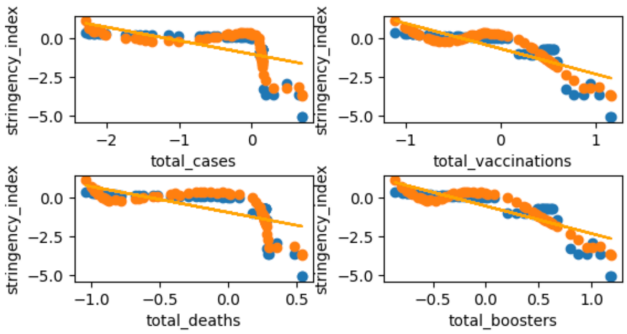

Figure 6. Colombia (Mid) Stringency Index vs. various independent variables after normalization. R^2 = 0.896, MSE = 0.222, RMSE = 0.471

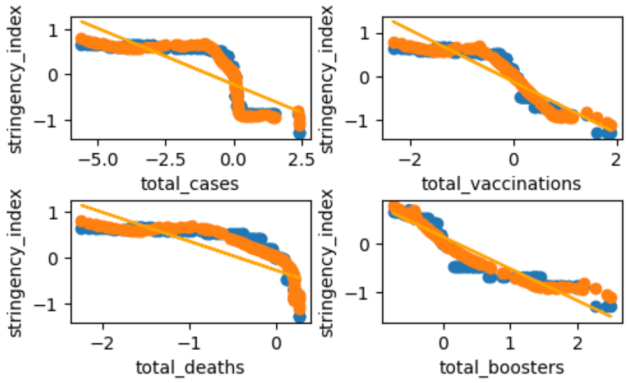

Figure 7. Cambodia (Low) Stringency Index vs. various independent variables after normalization. R^2 = 0.952, MSE = 0.0178, RMSE = 0.133

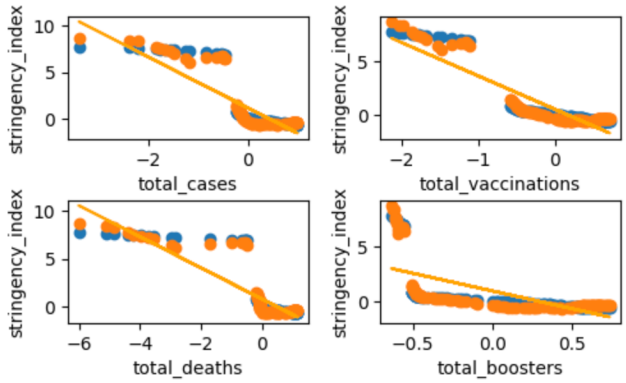

Figure 8. Pakistan (Low) Stringency Index vs. various independent variables after normalization. R^2 = 0.971, MSE = 0.138, RMSE = 0.372

Conclusion

The generated models provided high R-squared values with generally low MSE values indicating that it can explain the variation while also making predictions based on the trend between the correlation of the variables. It is, however, vital

to note that just because the trend exists doesn't mean that our independent variables are necessarily causing the dependent variable. Correlation between variables does not always imply causation. While we can make general comments on if one variable affects another, to understand the full cause of the behavior of a dependent variable, other variables need to be analyzed as they may be playing a role.

In order to construct the model with multiple variables, we left out a large amount of data entries that lacked the information we needed to construct a linear regression. Because of this some of the lower HDI countries have less data points to build the model with leading to less credibility in its patterns and trends set by our regression. It is important to know that the data is not perfect and is cumulated by the efforts of those who choose and decide to report these statistics despite the chaos of a pandemic. In order to improve the credibility of the model, it may be necessary to gather more data from the lower HDI countries or to use techniques such as imputation to fill in missing values. It is also important to keep in mind that any model is only as good as the data it is based on, and it is crucial to be aware of the limitations and potential biases of the data when interpreting the results of the model.

References

"Johns Hopkins Coronavirus Research Center." Johns Hopkins Coronavirus Resource Center, https://coronavirus.jhu.edu/map.html.

"Linear Regression." Yale University Department of Statistics and Data Science, Yale University, http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

Mathieu, Edouard, et al. "Coronavirus Pandemic (COVID-19)." Our World in Data, 5 Mar. 2020, https://ourworldindata.org/coronavirus.

"Machine Learning." IBM Cloud Learn Hub, IBM Cloud Education, https://www.ibm.com/cloud/learn/machine-learning.

"Ordinary Least Squares Linear Regression." Scikit-Learn, https://scikit-learn.org

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Russel, Stuart. Peter, Norvig. 2021. Chapter 19: Machine Learning from Examples. In *Artificial Intelligence A Modern Approach, Fourth Edition.* 145-232. Boston: Pearson.

Johns Hopkins University, "Data on COVID-19 (coronavirus)." *Our World in Data, 2020,* *https://github.com/owid/covid-19-data/blob/master/public/data/README.md*

"Data Preprocessing in Machine learning." JavaTpoint, 2021, *https://www.javatpoint.com/data-preprocessing-machine-learning*

Acknowledgements

We thank Sravya Kondrakunta for providing code support on this project.