**ANL488 FINAL PROJECT REPORT**

**EFFECT OF SOCIAL-ECONOMIC PARAMETERS ON LOGISTICS PERFORMANCE INDEX**



**Submitted by   
ALYSSA LAM JIA YU**

**SCHOOL OF BUSINESS  
Singapore University of   
Social Sciences**

**Presented to Singapore University of Social   
Sciences in partial fulfilment of the   
requirements for the   
Degree of Bachelor of Science in   
Business Analytics**

**2023**

# Abstract

This study presents a comprehensive analysis of using machine learning algorithms to predict the Logistics Performance Index (LPI) based on the social-economic parameters from the World Bank. It details the significance of logistics efficiency measurements and the importance of a wide range of social-economic parameters. As such, feature selection is important to pre-determine the important predictors and keep the dataset relevant.

Using partitioned datasets of 70:20:10 for training, testing and validating, the modelling algorithms CART, Linear, CHAID, Random Forest, and Neural Net are carried out. According to the modelling results, CART with boosting surpasses other models, emerging as the champion model with the most competitive performance measures.

The important predictors as a percentage of Gross Domestic Product (% of GDP) include Domestic general government health expenditure, Domestic credit to private sector, Domestic credit to private sector by banks, Agriculture, forestry, and fishing, value added, Services, value added, Research and development expenditure, and Broad money. These parameters serve as key indicators in accurately predicting LPI and offer valuable insights for decision-makers and stakeholders in the logistics industry.

Table of Contents

[Abstract i](#_Toc134391675)

[Chapter One Introduction 1](#_Toc134391676)

[CRISP-DM 3](#_Toc134391677)

[Phase 1: Business Understanding 3](#_Toc134391678)

[Chapter Two Literature Review 4](#_Toc134391679)

[Supervised Machine Learning in Logistics Performance 4](#_Toc134391680)

[Relationship between Logistics Performance Index and Global Competitiveness Index 5](#_Toc134391681)

[Predicting Logistics Performance Index with Gross Domestic Product 6](#_Toc134391682)

[Improving Logistics Performance Index with S&P Global Market Intelligence’s Economic Statistics 7](#_Toc134391683)

[Improving Logistics Performance Index with Global Competitiveness Index 9](#_Toc134391684)

[Chapter Three Data Understanding and Preparation 11](#_Toc134391685)

[Data Understanding and Preparation 11](#_Toc134391686)

[Phase 2: Data Understanding 11](#_Toc134391687)

[Phase 3: Data Preparation 15](#_Toc134391688)

[Chapter Four Modelling and Evaluation/Discussion 22](#_Toc134391689)

[Overview of Stream 22](#_Toc134391690)

[Phase 4: Modelling 22](#_Toc134391691)

[Phase 5: Evaluation 39](#_Toc134391692)

[Phase 6: Deployment 41](#_Toc134391693)

[Chapter Five Recommendations/Conclusion 42](#_Toc134391694)

[Recommendations 42](#_Toc134391695)

[Conclusion 42](#_Toc134391696)

[Future Work 43](#_Toc134391697)

[Personal Learning 43](#_Toc134391698)

[References 45](#_Toc134391699)

[Appendix A 48](#_Toc134391700)

[Appendix B 49](#_Toc134391701)

[Appendix C 50](#_Toc134391702)

# Chapter One Introduction

International trade relies heavily on logistics, which encompasses a range of functions from the transportation of goods, storage in warehouses, clearing goods at borders, and payment systems. Both private operators and goods owners play important roles as service providers in logistics, but public policy offices at the national, regional, and international levels also need to have a grasp on logistics performance at a country level to measure and evaluate trade and transport facilitation over time and across different countries (Ekici et al., 2016).

Logistics performance is an important factor in measuring the competitiveness of countries and their abilities to partake in international trade (Martí et al., 2014). The Logistics Performance Index (LPI) is a widely used tool for evaluating a country's logistics performance and benchmarking it against other countries. The association of trade and transport facilitation can be described using the LPI. It is published by the World Bank and is based on a survey of firms operating in the country, with quantitative and qualitative assessments (World Bank, n.d).

A composite indicator is created by combining multiple indicators into a single index, based on a model that captures the complex and multi-faceted concept being measured. It is a mathematical combination of sub-indicators used to measure concepts that cannot be represented by a single indicator alone. The composite indicator, LPI, is determined by examining six main components, using indicators such as customs, infrastructure, international shipments, logistics quality and competence, tracking and tracing, and timeliness. Their inclusion is based on a survey of logistics professionals. The LPI is then built using Principal Component Analysis (PCA) which is a statistical technique for dataset dimensionality reduction. The indicators have been combined and assigned appropriate weights, with scores ranging from 1 to 5, where 5 represents the best logistics performance (Martí et al., 2017).

Martí et al. (2017) suggested that policymakers can affect power to enhance a country's logistics performance by implementing policies that improve various aspects assessed by the LPI. This includes aspects such as streamlining customs and border clearance, upgrading trade and transport infrastructure, increasing access to affordable shipment options, enhancing the competence and quality of logistics services, improving the ability to monitor shipments, and ensuring shipments are delivered within the scheduled or expected time frame.

Arvis et al. (2014) showed that higher-income countries typically occupy the top 10 positions in the LPI, while lower-income countries tend to be ranked lower. The findings suggest that countries with top positions in the LPI tend to have better geographical area for logistics movement, while those in the bottom may face challenges such as conflicts weakening their development. At the same time, the gap between the values at the top narrows over time and could be due to infrastructure improvement to grow trade in low to middle-income countries.

In recent years, the study of logistics performance and its relationship with social-economic parameters has received increasing attention (Ekici et al., 2016). Research has shown that various factors, such as institutional quality (Rakauskienė & Petkevičiūtė-Stručko, 2022), government spending on infrastructure (Munim & Schramm, 2018), and depth of human capital (Munim & Schramm, 2018) can have a significant impact on a country's logistics performance.

In this study, the objectives are to predict the LPI with the World Bank Indicators as the social-economic parameters and to extract insights from the predictor importance to aid in decision-making by countries and businesses on international trade and allocation of investments.

Additionally, social-economic parameters from the World Bank suggested by Ekici et al. (2019) with reference to pillars from 2017-2018 Global Competitiveness Index (GCI) are explored and analysed against the LPI dataset. The GCI is published by the World Economic Forum. It is a useful tool for understanding the relationship between social-economic parameters and logistics performance, as it provides a comprehensive view of a country's competitiveness. It is calculated based on the average of the scores from the 12 pillars: institutions, infrastructure, macroeconomic stability, health and primary education, higher education and training, goods market efficiency, labour market efficiency, financial market development, technological readiness, market size, business sophistication, and innovation. Each of the pillars is comprised of several indicators that measure different aspects of the country’s competitiveness and is published annually by the World Economic Forum.

## CRISP-DM

In this study, CRISP-DM (Cross-Industry Standard Process for Data Mining) will be used as it is a widely accepted framework in the industry for conducting data mining and has been proven to be effective for data mining (IBM, 2021a). IBM SPSS Modeler will be used in the data mining process for clear and streamlined workflow (IBM, 2021b).

### Phase 1: Business Understanding

#### Business Problem

The business problem is the limited understanding of the relationship between social-economic parameters and logistics performance, particularly in the post-pandemic world where these parameters are dynamic. Understanding the impact of pre-COVID social-economic parameters on logistics performance is critical and serves as the foundation for a logistics sector pre-pandemic analysis. The competitiveness of a country in international trade is closely tied to its logistics performance, which can be evaluated using the Logistics Performance Index (LPI). Despite its widespread use, specific factors that impact a country's logistics performance, such as institutional quality, government investment in infrastructure, and human capital development can be further explored. This study aims to address this business problem by exploring the relationship between these factors and logistics performance, which may have important implications for businesses and governments in their decision-making on international trade and investment allocation.

#### Business Analytics Problem

The business analytics problem is to establish the effect of social-economic parameters on logistics performance using machine learning techniques. Patterns, correlations, and cause-and-effect relationships can provide insights into the factors that contribute to a country's logistics performance (Ekici et al.,2016; Ekici et al., 2019; Jomthanachai et al., 2022). This requires the collection and preparation of the data, followed by the application of data mining techniques such as regression analysis and predictive models. With the parameters identified by the champion model, businesses and governments can leverage on the results to do decision-making on international trade and investment allocation.

#### Data Mining Goal

There are many parameters and indicators available publicly, such as from the World Bank, Asian Development Bank, World Economic Forum etc. For consistency, indicators from the World Bank are used. The aim is to form a predictive model that has high accuracy with best performing error values in its prediction of LPI.

# Chapter Two Literature Review

## Supervised Machine Learning in Logistics Performance

Ekici et al. (2016)introduced Global Competitiveness Index (GCI) as the representative of social-economic parameters to relate logistics with global competitiveness using expert judgement for feature selection, and Artificial Neural Network (ANN) for prediction. Comparably, Jomthanachai et al. (2022)introduced economic and demographic statistics data from S&P Global Market Intelligence with machine learning algorithms to select features and concludes that Artificial Neural Network (ANN) is best effective in predicting the Logistics Performance Index (LPI). Ekici et al. (2019) evaluated GCI pillars relevance in improving logistics performance with a tree-augmented naïve Bayesian network (BN-TAN) and partial least square (PLS) path model, and importance-performance map analysis (IPMA).

## Relationship between Logistics Performance Index and Global Competitiveness Index

Ekici et al. (2016)looked into how competitiveness indicators impact each of the LPI indicators. It suggests that the improvement in a country’s logistics performance should be highly dependent on its competitiveness rating. It describes logistics to involve coordinating a series of activities and its performance is greatly impacted by government interventions such as building infrastructure, developing transport regulations, and implementing efficient customs procedures. These areas mentioned are closely related to the competitiveness of the country on the global stage. As such, the 12 pillars of Global Competitiveness Index are relevant to its study in understanding logistics performance better.

Data is extracted from two main sources in Ekici et al. (2016), namely, 2007, 2010 and 2013 editions of Global Competitiveness Index (GCI), World Economic Forum (WEF), and 2007, 2010 and 2014 editions of Logistics Performance Index (LPI), World Bank. Given the relevance of the 12 GCI pillars to the six main components of LPI, it invites subject-matter experts in transportation and logistics activities to review the 114 WEF indicators under the 12 GCI pillars and to only assign the related WEF indicators to the LPI indicators upon group consensus. Unrelated WEF indicators are not considered in the analysis.

Random data divisions were used for sampling. ANN, using MATLAB neural network toolbox, then trains and tests the WEF indicators in a 70:15:15 proportion for training, testing and validation datasets. Two layers of feed-forward networks are used. The first layer comprises of WEF indicators related to each LPI indicator and the second layer being the output layer comprises five output neurons representing Cumulative Belief Degrees (CBD) of the LPI indicators.

Using the trained ANN model on Turkey as part of its scenario analysis in a case study, Turkey’s WEF indicator values as inputs, and the average WEF indicator values from the top five ranking countries in the latest LPI as targets, the important WEF indicators are determined. The results of the scenario analysis show the top three important WEF indicators that interact the most with the respective LPI indicators. For example, to enhance “Customs” performance of Turkey, “Reliability of police services, “Favoritism in decisions of government officials” and “trade tariffs” should be attended to first. Similarly found in 2012 and 2014 editions of LPI reports, low-performing countries should pay more attention to “informal payments” which also refers to corrupt payment, if they plan to improve their logistics performance. In its conclusion, the scenario analysis finds that “Fixed broadband internet subscriptions” appeared three times as the most important WEF indicator across the six LPI indicators. For Turkey, to have an improvement in the LPI, the country has to put its focus on the enhancement of its broadband internet accessibility.

However, to consider a two-way interaction between logistics performance and social-economic parameters, Ekici et al. (2016)highlighted that it is possible for logistics improvement to have an impact on economic growth. To illustrate, new investments and updated infrastructure increase demands of goods and services. When travelling time decreases, producers can reach out to more distant markets and attract more foreign direct investment.

## Predicting Logistics Performance Index with Gross Domestic Product

Jayathilaka et al. (2022) conducted a study to find out how the logistics move the global trade and economy. Gross Domestic Product (GDP) is examined together with the LPI across continents. It uses Panel Regression Technique and Random Effect (RE) for modelling. Results conclude that LPI is positively related to net exports globally and particularly within Asia, Europe, and Oceania. While GDP is negatively related to net exports, particularly within Asia and African continent, it is positively related within Oceania and Middle East. The study also suggests that export to import is unidirectional causality.

## Improving Logistics Performance Index with S&P Global Market Intelligence’s Economic Statistics

To understand the economic factors behind the country’s logistics performance, Jomthanachai et al. (2022)conducted a range of machine learning techniques to determine the ideal set of economic attributes that best describes a variable to predict the respective LPI. Firstly, economic and demographic statistics data is obtained from S&P Global Market Intelligence. It comprises 52 features under five main categories. To match LPI’s editions, the data is narrowed down to five periods from 2009 to 2018, and only 26 features are selected, X1 to X26. This macroeconomic data is later divided into 70:30 for training and testing/validating model performance.

Jomthanachai et al. (2022)attempted to use two feature selection approaches, Correlation and Principal Component Analysis (PCA), and another integrated approach called Least Absolute Shrinkage and Selection Operator (LASSO) or Elastic-net (E-net). This part of the process is the data pre-processing phase.

Correlation study is performed on Microsoft Excel with its data analysis tool. A correlation matrix is formed and correlation value, R, is calculated for each set of input and output variables. 0 to 1 is positive correlation, where r value 0 to 0.25 is weak, 0.25 to 0.5 is fair, 0.5 to 0.75 is good and 0.75 to 1 is excellent. 0 to -1 is negative correlation, where r value 0 to -0.25 is weak, -0.25 to -0.5 is fair, -0.5 to -0.75 is good and -0.75 to -1 is excellent. 0 is nonlinear correlation. In its selection, the three predictor variables in set A have r values more than 0.5. For set B, another five variables are selected and included based on their strong or outstanding correlation (r > = 0.5) with any of set A’s variables.

PCA, using RStudio, generated four sets of features, sets C, D, and E based on the proportion of variance and set F using a PCA-biplot. Set C, D, and E are allocated as such; attributes detected in Principal Component (PC) 1 to PC3 have 46.33 percent variance fall under set c, likewise for PC1 to PC5 with 62.12 percent variance fall under set D and for PC1 to PC10 with 81.41 percent variance fall under set E. As for set F, a PCA biplot shows feature vectors close to the perpendicular to the LPI vector is strongly correlated and included.

LASSO and E-net, using RStudio, generated three sets of features, sets G, H, and I, with nine, 10 and 15 features respectively. LASSO selected nine features which also coincides with the variation of alpha value of 0.9 under E-net method. Other alpha values considered are 0.1, 0.25, 0.5 and 0.75. When alpha is 0.25, 0.5 and 0.75, 10 features are selected, forming set H. When alpha is 0.1, 15 features were selected and are non-shrink, contributing to set I.

Jomthanachai et al. (2022)explored the use of ANN, Multi-layer Perception ANN (MLP-ANN), Support Vector Regression (SVR), Random Forest Regression (RFR) and Ridge Regression for the modelling phase.

ANN and MLP-ANN are built on MATLAB 2020b Neural network toolbox, with the difference in the parameters of the hidden layer sizes, 1 x 10 (one hidden layer with ten nodes) for ANN and 10 x 10 (ten hidden layers with ten nodes each) for MLP-ANN. SVR uses RStudio with library packages *caret* and *e1071* for classification and regression training. RFR uses RStudio with library packages *caret* and *randomForest* for classification and regression. Penalised linear regression, or ridge regression, uses RStudio, *glmnet* package. It has an advantage in reducing the magnitude of the regression coefficient towards zero. It results in better generalisability for predicting unseen data. Though there is an obvious drawback in this model, overfitting, LASSO and E-net can help to alleviate this issue. Ridge regression is only used for regression.

The performance of the regression models is then evaluated using mean absolute errors (MSE), mean absolute percentage errors (MAPE), root-mean-square error (RMSE), Nash-Sutcliffe efficiency coefficient (NSE) and determination coefficient (R²).

In its analysis of the modelling results, all the criteria are narrowed down to the feature set, set C by PCA, and sets H and I by E-net to provide the best acceptable performance. With these three sets, a feature union and intersection operation are carried out. Further findings indicate that ANN is the most effective predictive model in the study.

The analysis of the modelling results shows that the most acceptable performance is achieved with the feature set C obtained through PCA and sets H and I obtained through Elastic Net. A feature union and intersection operation are performed on these three sets to obtain a more comprehensive feature set. The results of this analysis indicate that the Artificial Neural Network (ANN) is the most effective predictive model in the study.

## Improving Logistics Performance Index with Global Competitiveness Index

To add on, Ekici et al. (2019) aimed to improve logistics performance by identifying the GCI pillars that are significant in affecting logistic performance. Through the study, indirect effects can be identified from the relationships generated. Casual relationships are determined using the tree-augmented naïve Bayesian network (BN-TAN) model on “Netica” software, followed by partial least square (PLS) path model. The BN-TAN model is utilised initially to reduce the excessive potential outcomes of causal relationships among multiple variables. Subsequently, PLS is employed to determine significant relationships and perform statistical assessments on them. The process continues until a good fit model is established. One limitation of PLS is that it can be challenging to establish causal directions between variables due to a lack of domain knowledge, but this is addressed using the BN-TAN model. Finally, an importance-performance map analysis (IPMA) is employed to rank the priority of the factors based on their direct and indirect effects on the variables.

A total of 24 causal directions are established under BN-TAN model. Using SmartPLS software, all the 24 hypotheses are inserted into the PLS model. After some attempts, the final model with a good fit is formed. Bootstrapping is applied to test the significance of the hypothesised casual relationships. The study finds these pillars rank top 5 most important: Pillar 9 (Technological readiness), Pillar 5 (Higher educational and training), Pillar 12 (Innovation), Pillar 10 (Market size) and Pillar 2 (Infrastructure). Ekici et al. (2019) concluded with the possibility of a future work on using clustering of countries based on their LPI values and within each cluster, causal relationships among the WEF pillars and LPI values would be analysed. Policy makers would then be able to the results better with country-specific analysis on cluster-based BN models. It believes that LPI and GCI are bilaterally interrelated upon further studies.

Similarly, Ekici et al. (2016)suggested that improving the global competitiveness of the country may directly relate to the improvement of the indicators in LPI. It establishes the relationship between primary national competitiveness factors and a country’s logistics performance. A bilateral causal relationship is expected in logistics and trade as international trade demands transport and logistics growth, and logistics positively contributes to increased productivity, consumption and trade. Likewise, logistics and competitiveness both exist in a close and interdependent relationship.

As Ekici et al. (2016), Ekici et al. (2019) and Jomthanachai et al. (2022) found social-economic parameters from GCI to be closely related to LPI and Jayathilaka et al. (2022) found that GDP is correlated to LPI, so this study will use the social-economic parameters as a percentage of GDP to predict LPI.

# Chapter Three Data Understanding and Preparation

## Data Understanding and Preparation

### Phase 2: Data Understanding

#### Collecting Data

The Logistics Performance Index is a source from World Bank and it contains 21 topics and 1442 series covering 168 economics and countries across four income groups. The six years involved in this dataset are 2007, 2010, 2012, 2014, 2016 and 2018. The LPI values will be used as the Target later in the modelling. It was downloaded directly from “<https://lpi.worldbank.org/sites/default/files/International_LPI_from_2007_to_2018.xlsx>” with the file named *International\_LPI\_from\_2007\_to\_2018.xlsx* (World Bank, 2018). It contains six years of LPI values, specifically 2007, 2010, 2012, 2014, 2016 and 2018 (Figure 1).

Graphical user interface, application, table, Excel

Description automatically generated

Figure 1: Screenshot of the LPI and its indicators in *International\_LPI\_from\_2007\_to\_2018.xlsx*

The social-economic parameters datasets are downloadable directly on “*<http://databank.worldbank.org/data/download/WDI_excel.zip>*” in one zipped folder containing where all of the available datasets from 1960 to 2021 with more than 180 countries in *WDIEXCEL.xlsx* (World Bank, 2022). As this file contains all the possible World Bank datasets, the file is slightly more than 70 MB (Figure 2). It has data from 67 sources, that consists of 21 topics and 1442 series. It covers 266 economies and countries over seven income groups. A preliminary stage of data selection is shown in Figure 3. It is an essential tool for studies to gain insights into the global economic trends.

Graphical user interface, text

Description automatically generated

Figure 2: Screenshot of the indicators in *WDIEXCEL.xlsx*

Text

Description automatically generated with medium confidence

Figure 3: Screenshot for the 11 selected indicators for the study

#### Describing and Exploring Data

The number of countries surveyed for the formulation of LPI each year is shown in Table 2. To describe the LPI dataset, all the values of LPI over the six years are combined onto one spreadsheet. From the metadata, the income groups defined by World Bank are High Income, Upper Middle Income, Lower Middle Income and Lower Income. Statistics and normal distribution of LPI values can be observed in Figures 4 and 5.

Table 1: Overview of the number of countries in each income group from 2007 to 2018

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Income Groups** | **2007** | **2010** | **2012** | **2014** | **2016** | **2018** | **Total** |
| High income | 46 | 49 | 48 | 49 | 51 | 51 | 294 |
| Low income | 24 | 23 | 21 | 24 | 22 | 24 | 138 |
| Lower middle income | 46 | 41 | 44 | 44 | 45 | 44 | 264 |
| Upper middle income | 31 | 40 | 40 | 41 | 40 | 39 | 231 |
| **Total** | **147** | **153** | **153** | **158** | **158** | **158** | **927** |

Graphical user interface

Description automatically generated with low confidence Chart, histogram

Description automatically generated

Figure 4 and Figure 5: Statistics (left) and Normal Distribution (right) of LPI Values

With the given countries and income group status, exploratory visualisations on Tableau are created. According to the World Bank, on the one hand, low-income countries are predominantly located in Africa and some parts of South Asia. On the other hand, high-income countries are primarily found in North America, Europe, and Oceania. Lower middle-income countries are distributed along the coastal regions of Africa and South Asia to Southeast Asia. Upper middle-income countries are mainly located in South America, some parts of Africa, and Asia (Figure 7).

Year 2018 has a high concentration of LPI values under 3.0 with Afghanistan being the lowest ranked country at a score of 1.86 and Germany being the highest at a score of 4.18 (Figures 8 and 9). It can be observed that Germany had topped the LPI ranks for five out of the six available years while Afghanistan had been the lowest for twice (Figures 10 and 11).

Map

Description automatically generated

Figure 7: Visualisation of Countries in each Income Group on Tableau

Chart, histogram

Description automatically generated

Figure 8: LPI Distribution in 2018, highlighting Germany as the Best Performer

Chart, histogram

Description automatically generated

Figure 9: LPI distribution in 2018, highlighting Afghanistan as the Worst Performer

Table, Excel

Description automatically generated Table, Excel

Description automatically generated

Figure 10 and Figure 11: Best (left) and Worst (right) LPI Performers from 2007 to 2018

#### Verifying Data Quality

Even though datasets are downloaded on World Bank’s open data bank, the data needs to be verified and accessed for data accuracy, completeness, consistency, and relevance to the study topic.

### Phase 3: Data Preparation

#### Data Preparation using Microsoft Excel

The data needs to be cleaned and prepared for analysis. This may involve dealing with missing values, outliers, and other data quality issues. The original dataset *WDIEXCEL.xlsx* contains more than 380,000 rows of data comprising more than 260 countries/groups and 1442 indicators, and has data recorded from 1960 to 2021. To select the parameters for the study, indicators with “% of GDP” as part of the indicator name are used. There are 49 out of the 1442 indicators, and they belong to 17 out of 21 topics (Figures 12 and 13).

Since only 49 indicators from 2007 to 2021 are selected for the study, the file size is significantly reduced. However, the challenge is to flatten the data which may require the use of Python coding. To prepare the data, indicators related to LPI’s direct formulation are excluded (Figure 14). Names that are not specifically referring to the countries and territories are excluded.

Graphical user interface, text, application

Description automatically generated Text, email

Description automatically generated

Figure 12 and Figure 13: An extract showing a part of the selected Indicators (left) and Topics (right)

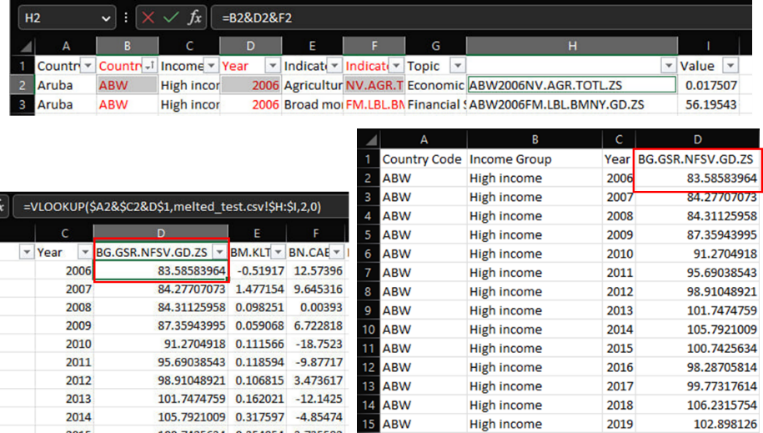


Figure 14: Preparing data to be readable on IBM SPSS Modeler

#### Data Preparation using IBM SPSS Modeler

The dataset is loaded in a comma separated values text file. Settings are adjusted in the file node to reflect delimiters to be commas (Figure 15). As the parameters are not read in as continuous values, overriding the storage type to “Real” is necessary so that these parameters can be in “Continuous” under Measurement (Figure 16). Country Code acts as Record ID, LPI score “LP.LPI.OVRL.XQ” acts as Target for the supervised machine learning, while the rest act as Inputs (Figure 17).

Graphical user interface, text, application, email

Description automatically generated

Figure 15: Settings on Input File Node

Graphical user interface, table

Description automatically generated

Figure 16: Assigning the type of data “Real” for numeric values for the Social-Economic Parameters

Table

Description automatically generated

Figure 17: Finalised settings of the Input file

#### Auto Data Prep Node

Auto Data Prep node removes fields with more than 50% missing records in a bid to achieve a balance between speed and accuracy for building models in the later steps (Figure 18). This step is to ensure adequate consideration with equal priority is given for modelling. To summarise, there were 1442 indicators initially. Since indicators with “% of GDP” were selected, there are 51 fields remaining which includes 49 indicators as Input variables, one Record ID variable and one Target variable. After applying Auto Data Prep node, there are 43 fields remaining which includes 41 indicators as Input variables, one Record ID variable and one Target variable. Figures 19, 20 and 21 show the results of the Auto Data Prep.

**Graphical user interface, text, application, email

Description automatically generated**

Figure 18: Settings on Auto Data Prep Node

**Chart

Description automatically generated**

Figure 19: Summary of Data Processing

**A picture containing diagram

Description automatically generated**

Figure 20: Field Details of LP.LPI.OVRL.XQ

Graphical user interface

Description automatically generated

Figure 21: Field Details of FS.AST.PRVT.GD.ZS

#### Feature Selection Process

Ekici et al. (2016) carried out feature selection in its study to select the best set of features to build models. Here, feature selection is used to allow a maximum 30% of missing values to increase the supervised machine learning models’ efficiencies (Figure 22). Table node shows 38 fields remaining (Figure 23) and Data Audit node shows 36 fields, with 34 Inputs, one Record ID and one Target (Figure 24).

Graphical user interface, text

Description automatically generated

Figure 22: Settings on Feature Selection Node

Table

Description automatically generated

Figure 23: Table Node on Feature Selection Results

A screenshot of a computer

Description automatically generated with medium confidence

Figure 24: Data Audit on Feature Selection Results

#### Partition for Training, Testing and Validation Datasets

To do the estimation using predictive models, a Partition node is applied to the Feature Selection nugget. The dataset is partitioned into 70% for training and 20% for testing datasets, and a 10% as unseen dataset for validation purpose (Figures 25 and 26). The purpose of the 70:20:10 partitioning allows the machine learning model performance to be evaluated on the testing dataset, and to assess the generalisation performance on new and unseen data on the validation dataset. This is a common practice in building robust and reliable models. Assigned labels are shown in Figure 27.

Graphical user interface, application

Description automatically generated Chart, pie chart

Description automatically generated

Figure 25 and Figure 26: Settings on Partition Node (70:20:10 Partitioning) (left) and Plot Visualisation (right)

Table

Description automatically generated

Figure 27: Assignment of Partition Labels on each record

# Chapter Four Modelling and Evaluation/Discussion

## Overview of Stream

Figure 28 shows the overview of the stream, with data preparation, modelling and analysis phases.

Chart

Description automatically generated

Figure 28: Overview of the Stream

### Phase 4: Modelling

The inputs and targets shown in Figure 29 are used on all the models in this study.

Graphical user interface, text, application

Description automatically generated

Figure 29: Predictors and Target Selection used in all models

#### CART Models

CART (Classification and Regression Trees) is a decision tree-based technique that can handle classification and regression problems (Chan et al., 2022). CART divides the dataset recursively into subgroups depending on the predictor variables, with binary splits. As the target is continuous, the splitting criterion is based on sum of squared errors, where a lower SSE means a smaller deviation between the target value and the mean of the particular node. CART is available in standard, boosting, and bagging forms, making it a versatile method for a variety of circumstances. CART can also work with a continuous (metric) target variable and is capable of handling high-dimensional datasets in general. Three CART nodes are connected to the Partition node to build standard, boosting and bagging CART models. The maximum tree depth is set to default, at 5. Applications of CART are from Figures 30 to 44.

Graphical user interface, application

Description automatically generated

Figure 30: Maximum tree depth for all CART models

Graphical user interface, text, application, email

Description automatically generated

Figure 31: CART Node Settings – Single (Standard), Boosting and Bagging Models

Graphical user interface, text, application, email

Description automatically generated

Figure 32: Component Models Settings for Boosting CART and Bagging CART

Table

Description automatically generated with medium confidence

Figure 33: Analysis Node on Single (Standard) CART

Table

Description automatically generated Table

Description automatically generated

Figure 34 and Figure 35: Analysis Node on Boosting CART (left) and Bagging CART (right)

Diagram, engineering drawing

Description automatically generated

Figure 36: Decision Tree of Single (Standard) CART with 16 leaf nodes

1Chart

Description automatically generated with medium confidence

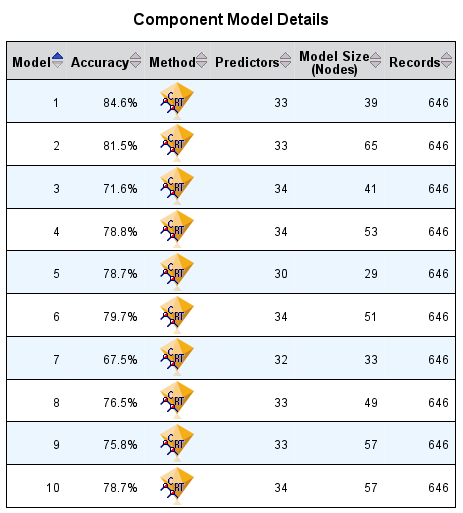
Figure 37: Decision Rules of Single (Standard) CART

Graphical user interface, chart

Description automatically generated with medium confidence Graphical user interface

Description automatically generated

Figure 38 and Figure 39: Model Summary of Boosting CART (left) and Bagging CART (right)

A screenshot of a graph

Description automatically generated with low confidence

Figure 40 and Figure 41: Component Model Details of Boosting CART (left) and Bagging CART (right)

Chart

Description automatically generated with medium confidence

Figure 42: Plot Node on Single (Standard) CART

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure 43 and Figure 44: Plot Node on Boosting CART (left) and Bagging CART (right)

#### Linear (Regression) Models Linear (Regression) Models

Linear regression is used to classify records based on continuous input types. This model involves fitting a straight line to the data to minimise the difference between the predicted and actual target values (Chan et al., 2022). It evaluates the predictor variable coefficients by minimizing the sum of squared errors between the actual and predicted target values. The linear model can be used in three ways: standard, boosting, and bagging. The standard version is the most basic linear regression implementation, whereas the boosting and bagging versions use ensemble techniques to create multiple models and aggregate their predictions to increase the model's overall accuracy (IBM, 2023). Applications of Linear are from Figures 45 to 57.

Graphical user interface, text, application, email

Description automatically generated

Figure 45: Linear (Regression) Node Settings – Standard, Boosting and Bagging Models

Graphical user interface

Description automatically generated

Figure 46: Model Summary of Linear (Regression) Standard Model

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 47 and Figure 48: Model Summary of Linear (Regression) Boosting Model (left) and Bagging Model (right)

A picture containing text, screenshot, number, font

Description automatically generatedA picture containing text, screenshot, number, font

Description automatically generated

Figure 49 and Figure 50: Component Model Details of Linear (Regression) Boosting Model (left) and Bagging Model (right)

Chart, sunburst chart

Description automatically generated

Figure 51: Coefficient Estimates by Linear (Regression) Standard Model

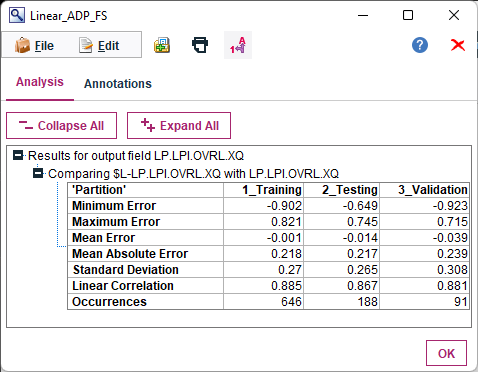


Figure 52: Analysis Node on Linear (Regression) Node Standard Model

Table

Description automatically generated Table

Description automatically generated

Figure 53 and Figure 54: Analysis Node on Linear (Regression) Boosting Model (left) and Bagging Model (right)

Chart, scatter chart

Description automatically generated

Figure 55: Plot Node on Linear (Regression) Standard Model

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 56 and Figure 57: Plot Node on Linear (Regression) Boosting Model (left) and Bagging Model (right)

#### CHAID Models

Similar to CART, CHAID (Chi-Squared Automatic-Interactions Detector) is a decision tree-based algorithm but uses the chi-squared test to identify the most significant predictor variables (Chan et al., 2022). Unlike CART, CHAID performs non-binary splits, making it more versatile for datasets with categorical variables. The algorithm recursively partitions the dataset into subsets based on the predictor variables to maximize the chi-squared test statistic. CHAID is also known for its ability to handle large datasets and to do multi-splits. CHAID is also able to create models that are standard, with boosting and with bagging. Applications of CHAID are from Figures 58 to 70.

Graphical user interface, text, application, email

Description automatically generated

Figure 58: CHAID Node Settings – Single (Standard), Boosting and Bagging Models

Table

Description automatically generated

Figure 59: Analysis Node on Single (Standard) CHAID

Table

Description automatically generated Table

Description automatically generated

Figure 60 and Figure 61: Analysis Node on Boosting CHAID (left) and Bagging CHAID (right)

A picture containing scatter chart

Description automatically generated

Figure 62: Decision Tree of Single (Standard) CHAID with 43 leaf nodes

Graphical user interface

Description automatically generated

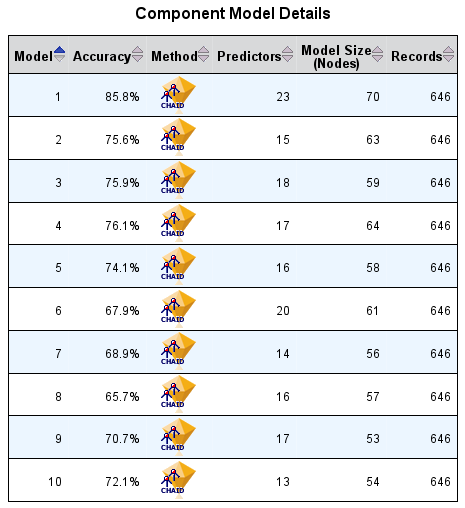
Figure 63: Decision Tree of Single (Standard) CHAID

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 64 and Figure 65: Model Summary of Boosting CHAID (left) and Bagging CHAID (right)

A screenshot of a graph

Description automatically generated with low confidence

Figure 66 and Figure 67: Component Model Details of Boosting CHAID (left) and Bagging CHAID (right)

Chart, scatter chart

Description automatically generated

Figure 68: Plot Node on Single (Standard) CHAID

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure 69 and Figure 70: Plot Node on Boosting CHAID (left) and Bagging CHAID

#### Random Forest Model

Random Forest is an ensemble algorithm itself that takes random observations by randomly sampling with replacements of the training dataset during the model building process (Chan et al., 2022). It also gets random inputs as part of the random sampling of input features subsets during the split of each node. With the two random schemes and the growing of a forest instead of a single tree, Random Forest is one of the most accurate models developed. Overfitting is unlikely to happen due to the control measures in place. Application of Random Forest is from Figures 71 to 74.

Graphical user interface, text, application, email

Description automatically generated

Figure 71: Random Forest Node Settings

Table

Description automatically generated

Figure 72: Analysis Node on Random Forest

A picture containing table

Description automatically generatedA picture containing table

Description automatically generatedChart

Description automatically generated with medium confidence

Figure 73: Predictor Importance of Random Forest

Chart, scatter chart

Description automatically generated

Figure 74: Plot Node on Boosting Random Forest

#### Neural Net Models

Neural Net uses one to multi-layer network of nodes (neurons) to predict the target. The model is made of an input layer, one or more hidden layers which is also known as the black box, and an output layer (Chan et al., 2022). The sample dataset undergoes the processing units of the network and has weights assigned to each connection during the learning process. The target is then returned by the output layer. The second layer of the two-layer neural net or the final layer of a multi-layer neural net will be responsible on combining all the signals from the preceding layer into one single output signal. Neural Net is useful in identifying highly complex patterns and predicting outcomes accurately. The downside would be the need of very large set of training dataset. Single (Standard), Boosting and Bagging versions are explored. Applications of Neural Net are from Figures 75 to 88.

Graphical user interface, text, application, email

Description automatically generated

Figure 75: Neural Net Node Settings – Single (Standard), Boosting and Bagging Models

Graphical user interface

Description automatically generated

Figure 76: Select Multilayer Perceptron on Neural Net Node Settings

Table

Description automatically generated

Figure 77: Analysis Node on Single (Standard) Neural Net

Table

Description automatically generated Table

Description automatically generated

Figure 78 and Figure 79: Analysis Node on Boosting Neural Net (left) and Bagging Neural Net (right)

A picture containing graphical user interface

Description automatically generated Diagram

Description automatically generated

Figure 80 and Figure 81: Model Summary of Single (Standard) Neural Net and Two-Layer Neural Network Diagram

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 82 and Figure 83: Model Summary of Boosting Neural Net (left) and Bagging Neural Net (right)

A screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

Description automatically generated with low confidence

Figure 84 and Figure 85: Component Model Details of Boosting Neural Net (left) and Bagging Neural Net (right)

Chart, scatter chart

Description automatically generated

Figure 86: Plot Node on Single (Standard) Neural Net

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure 87 and Figure 88: Plot Node on Boosting Neural Net (left) and Bagging Neural Net (right)

### Phase 5: Evaluation

In this phase, the models used in the modelling stage are assessed in terms of their performance to ensure that they are meeting the study’s objectives. The metrices to measure the machine learning models are the error values and the linear correlation values of the validation dataset which was partitioned earlier. The validation dataset is assigned to be 10% of the cleaned dataset. It is unseen and has not been used to train or test earlier so as to provide an unbiased assessment of the models’ generalisation capabilities. A good model will have low error values and high linear correlation value. Table 2 and 3 show the models in their standard, boosting and bagging versions. Representing models are indicated and to be compared across with all representatives. The champion model, Boosting CART, is one that can generalise better than the rest.

The performance measures for estimators with continuous target type (Chan et al., 2022; Jomthanachai et al., 2022) are:

1. Mean Absolute Error (MAE) for checking the absolute difference between the actual and predicted values , and divided by total count in the validation dataset .
2. Mean Absolute Percentage Error (MAPE) for a unit-free evaluation measure.
3. Mean Squared Error (MSE) for evaluating the average magnitude of squared errors,
4. Root Mean Squared Error (RMSE) with an alteration of y’s unit of measurement, and
5. Nash-Sutcliffe Efficiency Coefficient (NSE) where Is the average actual LPI value:
6. Determination Coefficient (R2) where Is the average predicted LPI value:

They seek to evaluate the results with average error and functions. Table 2 shows the overall analysis statistics of the models used. Table 2 shows the comparison of the performance measures and Boosting CART is found to be the champion model with the lowest error measurements.

Table 2: Comparison of Performance Measures 1 (Appendix A)



Table 3: Comparison of Performance Measures 2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MAPE** | **MSE** | **RMSE** | **NSE** | **R^2** | **R** | **Model Accuracy** |
| Standard CART | 0.22989 | 9.04197 | 0.09920 | 0.31497 | 0.78020 | 0.79006 | 0.88885 | - |
| **Boosting CART** | 0.19216 | 7.37707 | 0.05833 | 0.24152 | 0.86921 | 0.87220 | 0.93391 | 91.60% |
| Bagging CART | 0.22055 | 8.56075 | 0.08518 | 0.29186 | 0.80938 | 0.82482 | 0.90820 | 87.50% |
| Standard Linear | 0.69007 | 23.95666 | 0.70324 | 0.83860 | -0.71099 | 0.00174 | 0.04174 | 77.70% |
| Boosting Linear | 0.27015 | 10.40335 | 0.13229 | 0.36371 | 0.70982 | 0.69057 | 0.83101 | 77.90% |
| Bagging Linear | 0.24535 | 9.70002 | 0.09967 | 0.31571 | 0.79144 | 0.76595 | 0.87518 | 78.40% |
| Standard CHAID | 0.22507 | 8.63465 | 0.08603 | 0.29331 | 0.80714 | 0.80956 | 0.89976 | - |
| Boosting CHAID | 0.20995 | 7.90798 | 0.07449 | 0.27293 | 0.83295 | 0.83831 | 0.91559 | 90.40% |
| Bagging CHAID | 0.22487 | 8.57258 | 0.08829 | 0.29713 | 0.80236 | 0.82096 | 0.90607 | 89.40% |
| Random Forest | 0.19977 | 7.82509 | 0.06761 | 0.26002 | 0.84860 | 0.85931 | 0.92699 | - |
| Standard Neural Net | 0.23564 | 9.28143 | 0.10485 | 0.32381 | 0.76616 | 0.77053 | 0.87780 | 85.50% |
| Boosting Neural Net | 0.21213 | 8.07353 | 0.08091 | 0.28444 | 0.81854 | 0.82082 | 0.90599 | 99.00% |
| Bagging Neural Net | 0.27020 | 10.53850 | 0.13053 | 0.36129 | 0.70746 | 0.70954 | 0.84234 | 95.50% |
| **Best Performing Value** | 0.19216 | 7.37707 | 0.05833 | 0.24152 | 0.86921 | 0.87220 | 0.93391 |  |
| **Criteria** | Minimum Value | Minimum Value | Minimum Value | Minimum Value | Closest to 1 | Maximum Value | Maximum Value |  |
| **Champion Model** | Boosting CART | Boosting CART | Boosting CART | Boosting CART | Boosting CART | Boosting CART | Boosting CART |  |

### Phase 6: Deployment

The results of the analysis will be evaluated to determine the effect of social-economic parameters on LPI performance. Ekici et al., (2019) found the most important GCI pillars to be Pillar 9 (Technological readiness), Pillar 5 (Higher educational and training), Pillar 12 (Innovation), Pillar 10 (Market size) and Pillar 2 (Infrastructure) (Appendix B). This study finds the top predictors to be domestic health expenditure and domestic credit to private sector to be the most important, followed by the activities related to the cultivation of crops, livestock production, forestry, hunting and fishing, and service sector economic contribution (Table 4).

Table 6: Champion Model’s Boosting CART Predictor Importance (Appendix C)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ranking | Indicator Name | Indicator Code | Predictor Importance | Topic |
| 1 | Domestic general government health expenditure (% of GDP) | SH.XPD.GHED.GD.ZS | 0.11 | Health: Health systems |
| 2 | Domestic credit to private sector (% of GDP) | FS.AST.PRVT.GD.ZS | 0.11 | Financial Sector: Assets |
| 3 | Domestic credit to private sector by banks (% of GDP) | FD.AST.PRVT.GD.ZS | 0.11 | Financial Sector: Assets |
| 4 | Agriculture, forestry, and fishing, value added (% of GDP) | NV.AGR.TOTL.ZS | 0.07 | Economic Policy & Debt: National accounts: Shares of GDP & other  **GCI Pillar 10** |
| 5 | Services, value added (% of GDP) | NV.SRV.TOTL.ZS | 0.06 | Economic Policy & Debt: National accounts: Shares of GDP & other  **GCI Pillar 10** |
| 6 | Research and development expenditure (% of GDP) | GB.XPD.RSDV.GD.ZS | 0.06 | Infrastructure: Technology  **GCI Pillar 9 and Pillar 12** |
| 7 | Broad money (% of GDP) | FM.LBL.BMNY.GD.ZS | 0.04 | Financial Sector: Monetary holdings (liabilities) |
| 8 | Current health expenditure (% of GDP) | SH.XPD.CHEX.GD.ZS | 0.04 | Health: Health systems |
| 9 | Households and NPISHs final consumption expenditure (% of GDP) | NE.CON.PRVT.ZS | 0.03 | Economic Policy & Debt: National accounts: Shares of GDP & other  **GCI Pillar 10** |
| 10 | External balance on goods and services (% of GDP) | NE.RSB.GNFS.ZS | 0.03 | Economic Policy & Debt: National accounts: Shares of GDP & other |

# Chapter Five Recommendations/Conclusion

## Recommendations

The results of the study will be presented to stakeholders, including businesses and governments, to aid in their decision-making on international trade and allocation of investments. The identification of the important predictors of LPI generated by Boosting CART may also be used to guide future research in the field of logistics performance.

By addressing the business problem through the CRISP-DM framework, this study uncovers the relationship between social-economic parameters and logistics performance, which can have important implications for businesses and governments, and highlights and explains the significance of each selected parameter for LPI improvement.

## Conclusion

This study explores the various machine learning methods; CART, Linear, CHAID, Random Forest and Neural Net models. Boosting and Bagging options were applied to reduce the errors in CART, Linear, CHAID and Neural Net models. The champion model was found to be CART with Boosting. Boosting allowed the machine learning model to reduce variance and bias in an ensemble. It also handles overfitting easily comparatively. As boosting is sensitive to outliers, the dataset has been prepared to exclude outliers using Auto Data Prep on IBM SPSS Modeler.

## Future Work

As LPI 2023 report and dataset have been released in April 2023, future work can focus on understanding the impact of the COVID-19 pandemic on LPI. It would be interesting to find out and assess the updated important predictors of LPI given the pandemic situation.

## Personal Learning

This course has provided me with the opportunity to grow and expand my knowledge by pushing me beyond my comfort zone and triggering my curiosity in machine learning. Through engaging in research and readings on topics related to logistics performance and machine learning, I have been able to gain a more comprehensive understanding of the topics. The work in this project has been done iteratively so that insights and findings are well-aligned with the business objectives, which are also in line with the CRISP-DM methodology. At one time, I was attempting to use Python to code all the way but decided to continue with IBM SPSS Modeler due to my familiarity with the software. While Python offers a versatile and powerful platform for the data mining project, I have included its use in the data collection and preparation stages. The IBM SPSS Modeler was preferred for modelling as it comes with built-in tools, algorithms, and models, and I could work with various models effectively and efficiently given the timeline and workload on hand. This invaluable experience has provided me with valuable skills and knowledge that will be instrumental in my future work.

(Word Count from Chapter One to Chapter Five: 6840)

# References

Arvis, J.-F., Saslavsky, D., Ojala, L., Shepherd, B., Busch, C., & Raj, A. (2014). *Connecting to Compete*. Retrieved February 13, 2023, from <https://www.worldbank.org/content/dam/Worldbank/document/Trade/LPI2014.pdf>

Chan, S. P., Tan, S. C., Wang, D., & Ren, J. (2022). *ANL307 Predictive modelling (study*

*guide)*. Singapore: Singapore University of Social Sciences.

IBM. (2021). CRISP-DM Help Overview. Retrieved May 7, 2023, from <https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview>

IBM. (2021b). CRISP-DM in IBM SPSS Modeler. Retrieved May 7, 2023, from <https://www.ibm.com/docs/en/spss-modeler/saas?topic=overview-crisp-dm-in-spss-modeler>

IBM. (2023). *Linear Node*. Linear node (SPSS Modeler) - IBM Cloud Pak for Data as a Service. Retrieved May 7, 2023, from <https://dataplatform.cloud.ibm.com/docs/content/wsd/nodes/linear.html>

Jayathilaka, R., Jayawardhana, C., Embogama, N., Jayasooriya, S., Karunarathna, N., Gamage, T., & Kuruppu, N. (2022). Gross domestic product and logistics performance index drive the world trade: A study based on all continents. *PLOS ONE*, *17*(3). <https://doi.org/10.1371/journal.pone.0264474>

Jomthanachai, S., Wong, W. P., & Khaw, K. W. (2022). An application of machine learning regression to feature selection: A Study of Logistics Performance and economic attribute. *Neural Computing and Applications*, *34*(18), 15781–15805. <https://doi.org/10.1007/s00521-022-07266-6>

Martí, L., Martín, J. C., & Puertas, R. (2017). A DEA-Logistics Performance index. *Journal of Applied Economics*, *20*(1), 169–192. <https://doi.org/10.1016/s1514-0326(17)30008-9>

Martí, L., Puertas, R., & García, L. (2014). The importance of the logistics performance index in International Trade. *Applied Economics*, *46*(24), 2982–2992. <https://doi.org/10.1080/00036846.2014.916394>

Munim, Z. H., & Schramm, H.-J. (2018). The impacts of port infrastructure and logistics performance on economic growth: The mediating role of Seaborne Trade. *Journal of Shipping and Trade*, *3*(1). <https://doi.org/10.1186/s41072-018-0027-0>

Önsel Ekici, Ş., Kabak, Ö., & Ülengin, F. (2016). Linking to compete: Logistics and global competitiveness interaction. *Transport Policy*, *48*, 117–128. <https://doi.org/10.1016/j.tranpol.2016.01.015>

Önsel Ekici, Ş., Kabak, Ö., & Ülengin, F. (2019). Improving logistics performance by reforming the Pillars of Global Competitiveness Index. *Transport Policy*, *81*, 197–207. <https://doi.org/10.1016/j.tranpol.2019.06.014>

Rakauskienė, O. G., & Petkevičiūtė-Stručko, M. (2022). Determinants of Logistics’ performance: A new approach towards analysis of economic corridors and Institutional Quality Impact. *Insights into Regional Development*, *4*(3), 11–33. <https://doi.org/10.9770/ird.2022.4.3(1)>

World Bank. (n.d.). *About*. About | Logistics Performance Index. Retrieved February 13, 2023, from <https://lpi.worldbank.org/about>

World Bank. (2018). *Home*. Home | Logistics Performance Index. Retrieved February 13, 2023, from <https://lpi.worldbank.org/>

World Bank. (2022). World Development Indicators. Retrieved February 13, 2023, from <https://datatopics.worldbank.org/world-development-indicators/>

# Appendix A

Table A1: Model Analyses Statistics



# Appendix B

Table B1: Pillars of Global Competitiveness Index 2017-2018 (Ekici et al., 2019)

|  |  |  |
| --- | --- | --- |
| Global Competitiveness Index 2017-2018 | | Ranking |
| Pillar 1 | Institutions |  |
| Pillar 2 | Infrastructure | 5 |
| Pillar 3 | Macroeconomic environment |  |
| Pillar 4 | Health and primary education |  |
| Pillar 5 | Higher education and training | 2 |
| Pillar 6 | Goods market efficiency |  |
| Pillar 7 | Labour market efficiency |  |
| Pillar 8 | Financial market development |  |
| Pillar 9 | Technological readiness | 1 |
| Pillar 10 | Market size | 4 |
| Pillar 11 | Business sophistication |  |
| Pillar 12 | Innovation | 3 |

# Appendix C

Table C1: Codes and Descriptions

|  |  |
| --- | --- |
| **Code** | **Description** |
| LP.LPI.OVRL.XQ | Logistics Performance Index: Overall (1=low to 5=high) |
| BG.GSR.NFSV.GD.ZS | Trade in services (% of GDP) |
| BM.KLT.DINV.WD.GD.ZS | Foreign direct investment, net outflows (% of GDP) |
| BN.CAB.XOKA.GD.ZS | Current account balance (% of GDP) |
| BX.KLT.DINV.WD.GD.ZS | Foreign direct investment, net inflows (% of GDP) |
| BX.TRF.PWKR.DT.GD.ZS | Personal remittances, received (% of GDP) |
| CM.MKT.LCAP.GD.ZS | Market capitalization of listed domestic companies (% of GDP) |
| CM.MKT.TRAD.GD.ZS | Stocks traded, total value (% of GDP) |
| FD.AST.PRVT.GD.ZS | Domestic credit to private sector (% of GDP) |
| FM.LBL.BMNY.GD.ZS | Broad money (% of GDP) |
| FS.AST.DOMO.GD.ZS | Domestic credit provided by financial sector (% of GDP) |
| FS.AST.DOMS.GD.ZS | Domestic credit to private sector by banks (% of GDP) |
| FS.AST.PRVT.GD.ZS | Private credit bureau coverage (% of adults) |
| GB.XPD.RSDV.GD.ZS | Research and development expenditure (% of GDP) |
| GC.AST.TOTL.GD.ZS | Central government debt, total (% of GDP) |
| GC.DOD.TOTL.GD.ZS | Central government debt, total (% of GDP) |
| GC.LBL.TOTL.GD.ZS | Claims on central government, etc. liabilities (% of GDP) |
| GC.NFN.TOTL.GD.ZS | Net foreign assets (% of GDP) |
| GC.NLD.TOTL.GD.ZS | Net domestic credit (% of GDP) |
| GC.REV.XGRT.GD.ZS | Revenue, excluding grants (% of GDP) |
| GC.TAX.TOTL.GD.ZS | Tax revenue (% of GDP) |
| GC.XPN.TOTL.GD.ZS | Expense (% of GDP) |
| MS.MIL.XPND.GD.ZS | Military expenditure (% of GDP) |
| NE.CON.GOVT.ZS | General government final consumption expenditure (% of GDP) |
| NE.CON.PRVT.ZS | Household final consumption expenditure, etc. (% of GDP) |
| NE.CON.TOTL.ZS | Final consumption expenditure (% of GDP) |
| NE.DAB.TOTL.ZS | Gross domestic savings (% of GDP) |
| NE.EXP.GNFS.ZS | Exports of goods and services (% of GDP) |
| NE.GDI.FPRV.ZS | Gross fixed capital formation, private sector (% of GDP) |
| NE.GDI.FTOT.ZS | Gross fixed capital formation (% of GDP) |
| NE.GDI.TOTL.ZS | Gross capital formation (% of GDP) |
| NE.IMP.GNFS.ZS | Imports of goods and services (% of GDP) |
| NE.RSB.GNFS.ZS | External balance on goods and services (% of GDP) |
| NE.TRD.GNFS.ZS | Trade (% of GDP) |
| NV.AGR.TOTL.ZS | Agriculture, forestry, and fishing, value added (% of GDP) |
| NV.IND.MANF.ZS | Manufacturing, value added (% of GDP) |
| NV.IND.TOTL.ZS | Industry (including construction), value added (% of GDP) |
| NV.SRV.TOTL.ZS | Services, value added (% of GDP) |
| NY.GDP.COAL.RT.ZS | Coal rents (% of GDP) |
| NY.GDP.FRST.RT.ZS | Forest rents (% of GDP) |
| NY.GDP.MINR.RT.ZS | Mineral rents (% of GDP) |
| NY.GDP.NGAS.RT.ZS | Natural gas rents (% of GDP) |
| NY.GDP.PETR.RT.ZS | Oil rents (% of GDP) |
| NY.GDP.TOTL.RT.ZS | Total natural resources rents (% of GDP) |
| NY.GDS.TOTL.ZS | Gross domestic savings (% of GDP) |
| NY.GNS.ICTR.ZS | Gross national savings (% of GDP) |
| SE.XPD.TOTL.GD.ZS | Government expenditure on education, total (% of GDP) |
| SH.XPD.CHEX.GD.ZS | Current health expenditure (% of GDP) |
| SH.XPD.GHED.GD.ZS | Domestic general government health expenditure (% of GDP) |
| TG.VAL.TOTL.GD.ZS | Merchandise trade (% of GDP) |