Crop Products Export Value Forecasting in Ethiopia

934G5 | Machine Learning

UNIVERSITY
OF SUSSEX

Candidate Number: [redacted]

Contents:

1. Introduction

This project aims to develop a multilayer perceptron (MLP) model for predicting the future export value of crop products in Ethiopia and other countries. The client intends to utilise the predicted crop products' export value to formulate an export strategy aimed at boosting Ethiopia's economy. At the same time, the client wants to utilise this model for predictions for other countries. Consequently, they require an MLP model to conduct forecasting of not only Ethiopia's export value but also other regions' value. The forecasted export values should represent the values three years into the future from the year of the independent variables, as specified by the client's guidelines.

2. Data Preparation

FAOSTAT has been used to collect information related to agriculture and food, especially focusing on cereals. According to Appendix 1, 13 different data sources, namely consumer price indicators, crop production indicators, exchange rates and so on, have been investigated. The data sources are collected on both a monthly and annual basis. However, this project intends to use only year-based data. Consequently, if data is available on a monthly basis such as the food index, food price inflation index, and exchange rate, the data will be replaced as its mean of a single year.

Equation 1. Monthly to Yearly

$$\mathbf{y} = \frac{\mathbf{m_1} + \mathbf{m_2} + \mathbf{m_3} + \cdots + \mathbf{m_{10}} + \mathbf{m_{11}} + \mathbf{m_{12}}}{\mathbf{12}}$$

Where m denotes monthly basis data and y represents year data

Since the data team cannot ascertain whether these variables are unrelated to the dependent variable, it would be preferable to substitute them with the average values for each year at least.

The datasets were merged by combining the data for each country and year. In many cases, this resulted in over 4,000 data points for each combination of country and year. To ensure data balance, variables with fewer than 2,000 data points in total were excluded from the dataset. For instance, the fertiliser usage dataset was omitted due to a significant number of missing values, resulting in all variables having less than 2,000 data points. In addition, for indicators composed of a three-year aggregate, such as the average dietary energy supply adequacy, the year data in the dataset has been updated based on the last year and amended accordingly. Furthermore, in the emissions dataset, the combination of element code F1712 and item code 7230 duplicates

the information found with the combination of element code F1712 and item code 72440. To address similar issues across the entire dataset, redundant values are resolved by selecting only one variable. Last but not least, it is worth noting that some of the data are estimates, while others are official figures. However, this project treats them all as meaningful data and includes them without distinction in order to secure as much data as possible.

As a result, a total of 100 independent variables were gathered. These variables were selected due to satisfying several conditions ① having more than 2,000 data points ② appear to have direct or indirect associations with crop products ③ No duplicate information with other variables. For more details on the independent variables, please refer to Table 1.

Table 1. List of Independent Variable

Name	Information	Unit	Note
Area	Countries		
Year	Years		
Forecast_Year	Year - three years into the future		
Value_23013	Consumer Prices (Food index)		Average (Months to Year)
Value_23014	Food Price Inflation	%	Average (Months to Year)
cpi_F1717	Yield of Cereals	100g/ha	
cpi_F1804	Yield of Citrus Fruit	100g/ha	
cpi_F17530	Yield of Fibre Crops	100g/ha	
cpi_F1738	Yield of Fruit	100g/ha	
cpi_F1841	Yield of Oilcrops (Cake Equivalent)	100g/ha	
cpi_F1732	Yield of Oilcrops (Oil Equivalent)	100g/ha	
cpi_F1726	Yield of Pulses	100g/ha	
cpi_F1720	Yield of Roots and Tubers	100g/ha	
cpi_F1723	Yield of Sugar Crops	100g/ha	
cpi_F1729	Yield of Treenuts	100g/ha	
cpi_F1735	Yield of Vegetables	100g/ha	
Value_emi_F1712_72430	N2O Emissions from the all crops	kt	
Value_emi_F1712_72440	CH4 Emissions from the all crops	kt	
Value_emi_6727_7230	N2O Emissions from the cropland organic soils	kt	
Value_emi_6728_7230	N2O Emissions from the grassland organic soils	kt	
Value_emi_6727_7273	CO2 Emissions from the cropland organic soils	kt	
Value_emi_6728_7273	CO2 Emissions from the grassland organic soils	kt	
Value_employ_21144	Employment in agriculture, forestry and fishing	1000 No	

exchange_rate	Exchange rate 1 USD to each currency		Average (Months to Year)
Value_fbi_S2905_5611	Import Quantity of Cereals (Excluding Beer)	1000t	
Value_fbi_S2907_5611	Import Quantity of Starchy Roots	1000t	
Value_fbi_S2909_5611	Import Quantity of Sugar & Sweeteners	1000t	
Value_fbi_S2911_5611	Import Quantity of Pulses	1000t	
Value_fbi_S2912_5611	Import Quantity of Treenuts	1000t	
Value_fbi_S2913_5611	Import Quantity of Oilcrops	1000t	
Value_fbi_S2914_5611	Import Quantity of Vegetable Oils	1000t	
Value_fbi_S2918_5611	Import Quantity of Vegetables	1000t	
Value_fbi_S2919_5611	Import Quantity of Fruits (Excluding Wine)	1000t	
Value_fbi_S2922_5611	Import Quantity of Stimulants	1000t	
Value_fbi_S2923_5611	Import Quantity of Spices	1000t	
Value_fbi_S2905_5911	Export Quantity of Cereals (Excluding Beer)	1000t	
Value_fbi_S2907_5911	Export Quantity of Starchy Roots	1000t	
Value_fbi_S2909_5911	Export Quantity of Sugar & Sweeteners	1000t	
Value_fbi_S2913_5911	Export Quantity of Oilcrops	1000t	
Value_fbi_S2914_5911	Export Quantity of Vegetable Oils	1000t	
Value_fbi_S2918_5911	Export Quantity of Vegetables	1000t	
Value_fbi_S2919_5911	Export Quantity of Fruits (Excluding Wine)	1000t	
Value_fbi_S2922_5911	Export Quantity of Stimulants	1000t	
Value_fbi_S2923_5911	Export Quantity of Spices	1000t	
Value_fbi_S2905_5123	Losses of Cereals (Excluding Beer)	1000t	
Value_fbi_S2907_5123	Losses of Starchy Roots	1000t	
Value_fbi_S2911_5123	Losses of Pulses	1000t	
Value_fbi_S2913_5123	Losses of Oilcrops	1000t	
Value_fbi_S2918_5123	Losses of Vegetables	1000t	
Value_fbi_S2919_5123	Losses of Fruits (Excluding Wine)	1000t	
Value_fbi_S2914_5154	Other uses of Vegetable Oils	1000t	
Value_fbi_S2905_5142	Food of Cereals (Excluding Beer)	1000t	
Value_fbi_S2907_5142	Food of Starchy Roots	1000t	
Value_fbi_S2909_5142	Food of Sugar & Sweeteners	1000t	
Value_fbi_S2911_5142	Food of Pulses	1000t	
Value_fbi_S2912_5142	Food of Treenuts	1000t	
Value_fbi_S2913_5142	Food of Oilcrops	1000t	
Value_fbi_S2914_5142	Food of Vegetable Oils	1000t	
Value_fbi_S2918_5142	Food of Vegetables	1000t	
Value_fbi_S2919_5142	Food of Fruits (Excluding Wine)	1000t	
Value_fbi_S2922_5142	Food of Stimulants	1000t	

Value_fbi_S2923_5142	Food of Spices	1000t	
Value_fsi21010	Average dietary energy supply adequacy	%	3-year average
Value_fsi21013	Average protein supply	g/cap/day	3-year average
Value_fsi21035	Cereal import dependency ratio	%	3-year average
Value_fsi21034	Percent of arable land equipped for irrigation	%	3-year average
Value_fsi21033	Value of food imports in total merchandise exports	%	3-year average
Value_fsi21032	Political stability and absence of violence/terrorism	Index	
Value_fsi21030	Per capita food production variability	1000 I$	
Value_fsi21031	Per capita food supply variability	kcal/cap/day	
Value_fsi21043	Prevalence of anaemia among women of reproductive age	%	
Value_fsi21049	Prevalence of low birthweight	%	
Export_Value	Total Export Value of Crop Products	1000 USD	
Import_Value	Total Import Value of Crop Products	1000 USD	
New_Import_Value	Total Import Value of Crop Products	1000 USD	3 Years into the future
Value_fdi_23082	Total FDI Inflows	USD	
Value_fdi_23085	Total FDI Outflows	USD	
Value_ltc_7271	Land Temperature Change	°c	
Value_ltc_6078	Land Temperature Change (STD)	°c	
Value_lu_6600	Land Usage of Country area	1000ha	
Value_lu_6601	Land Usage of Land area	1000ha	
Value_lu_6602	Land Usage of Agriculture	1000ha	
Value_lu_6610	Land Usage of Agricultural land	1000ha	
Value_lu_6620	Land Usage of Cropland	1000ha	
Value_lu_6621	Land Usage of Arable land	1000ha	
Value_lu_6630	Land Usage of Temporary crops	1000ha	
Value_lu_6633	Land Usage of Temporary meadows and pastures	1000ha	
Value_lu_6640	Land Usage of Temporary fallow	1000ha	
Value_lu_6650	Land Usage of Permanent crops	1000ha	
Value_lu_6655	Land Usage of Permanent meadows and pastures	1000ha	
Value_lu_6690	Land Usage of Land area equipped for irrigation	1000ha	
Value_pu_5157_1357	Agricultural Use - Pesticides (total)	t	
Value_pu_5157_1309	Agricultural Use - Insecticides	t	
Value_pu_5157_1320	Agricultural Use - Herbicides	t	
Value_pu_5157_1331	Agricultural Use - Fungicides and Bactericides	t	
Value_pu_5157_1352	Agricultural Use – Fungicides (Seed treatments)	t	
Value_pu_5157_1353	Agricultural Use - Insecticides (Seed Treatments)	t	
Value_pu_5157_1345	Agricultural Use - Rodenticides	t	
Value_pu_5159_1357	Use per area of cropland - Pesticides (total)	Kg/ha	

Value_pu_5173_1357	Use per value of agricultural production - Pesticides (total)	g/int$	

The export value from the food trade indicators dataset has been selected as the dependent variable. The data team excluded items such as other food, meat and meat preparations, alcoholic beverages, dairy products and eggs, tobacco, non-alcoholic beverages, and non-edible fats and oils from the dataset, focusing solely on crop products. Then, the total export value by summing up the export values of all crop items, namely cereals, sugar and honey, and fruit and vegetables, for each region and year was calculated and stored in the 'Export_Value' variable through equation 2.

Equation 2. Export Value of Crop Products

$$\textbf{Export Value} = c_1 + c_2 + \cdots + c_{n-1} + c_{n+1}$$

Where c denotes the crop products

Subsequently, the original export values were substituted with the export values three years into the future since this model is going to predict the future export value. In cases where future values were not available for certain years, they were replaced with a missing value marker from Numpy, denoted as np.nan. These new export values have been saved to the 'New_Export_Value' column and will be used as a dependent variable. In contrast, the original export values were not dropped from the dataset to use as independent variables. The 'Import_Value' variable was also computed similarly to the 'Export_Value' and saved in the variable 'New_Import_Value'.

Last but not least, this dataset is composed of information that is permitted for public use and collected through lawful procedures that do not violate data ethics. Moreover, the dataset will not be utilised for any purposes other than its intended goal.

3. Data Processing

3-1. Data Preprocessing

First of all, the number of missing values for each independent variable was calculated. As a result, 62 out of 100 variables were removed due to having more missing values than 974 data points which is approximately 20% of the total data points, 4,872. It is important to note that all missing values in 'New_Import_Value' appear as missing values in the same rows of the dependent variable 'New_Export_Value', they were retained at that stage of missing value removal since they would

be removed in subsequent steps.

- Deleted Variables' Name: 'cpi_F1804', 'cpi_F17530', 'cpi_F1841', 'cpi_F1726', 'cpi_F1723', 'cpi_F1729', 'Value_fbi_S2905_5611', 'Value_fbi_S2907_5611', 'Value_fbi_S2909_5611', 'Value_fbi_S2911_5611', 'Value_fbi_S2912_5611', 'Value_fbi_S2913_5611', 'Value_fbi_S2914_5611', 'Value_fbi_S2918_5611', 'Value_fbi_S2919_5611', 'Value_fbi_S2922_5611', 'Value_fbi_S2923_5611', 'Value_fbi_S2905_5911', 'Value_fbi_S2907_5911', 'Value_fbi_S2909_5911', 'Value_fbi_S2913_5911', 'Value_fbi_S2914_5911', 'Value_fbi_S2918_5911', 'Value_fbi_S2919_5911', 'Value_fbi_S2922_5911', 'Value_fbi_S2923_5911', 'Value_fbi_S2905_5123', 'Value_fbi_S2907_5123', 'Value_fbi_S2911_5123', 'Value_fbi_S2913_5123', 'Value_fbi_S2918_5123', 'Value_fbi_S2919_5123', 'Value_fbi_S2914_5154', 'Value_fbi_S2905_5142', 'Value_fbi_S2907_5142', 'Value_fbi_S2909_5142', 'Value_fbi_S2911_5142', 'Value_fbi_S2912_5142', 'Value_fbi_S2913_5142', 'Value_fbi_S2914_5142', 'Value_fbi_S2918_5142', 'Value_fbi_S2919_5142', 'Value_fbi_S2922_5142', 'Value_fbi_S2923_5142', 'Value_fsi21010', 'Value_fsi21013', 'Value_fsi21035', 'Value_fsi21034', 'Value_fsi21033', 'Value_fsi21032', 'Value_fsi21030', 'Value_fsi21031', 'Value_fsi21043', 'Value_fsi21049', 'Value_fdi_23085', 'Value_ltc_6078', 'Value_lu_6690', 'Value_pu_5157_1352', 'Value_pu_5157_1353', 'Value_pu_5157_1345', 'Value_pu_5159_1357', 'Value_pu_5173_1357'

When the variables were extracted from FAOSTAT, the data team included all relevant variables that have more than 2,000 data points because some variables with insufficient data were intended to be removed in this data preprocessing stage one more time. Afterwards, the entire missing values were detected and deleted from the dataset, which resulted in the elimination of 2,359 data points. Even if these were a huge amount of data points considering the total data points of the dataset, the data team decided to exclude them due to the absence of the data collector.

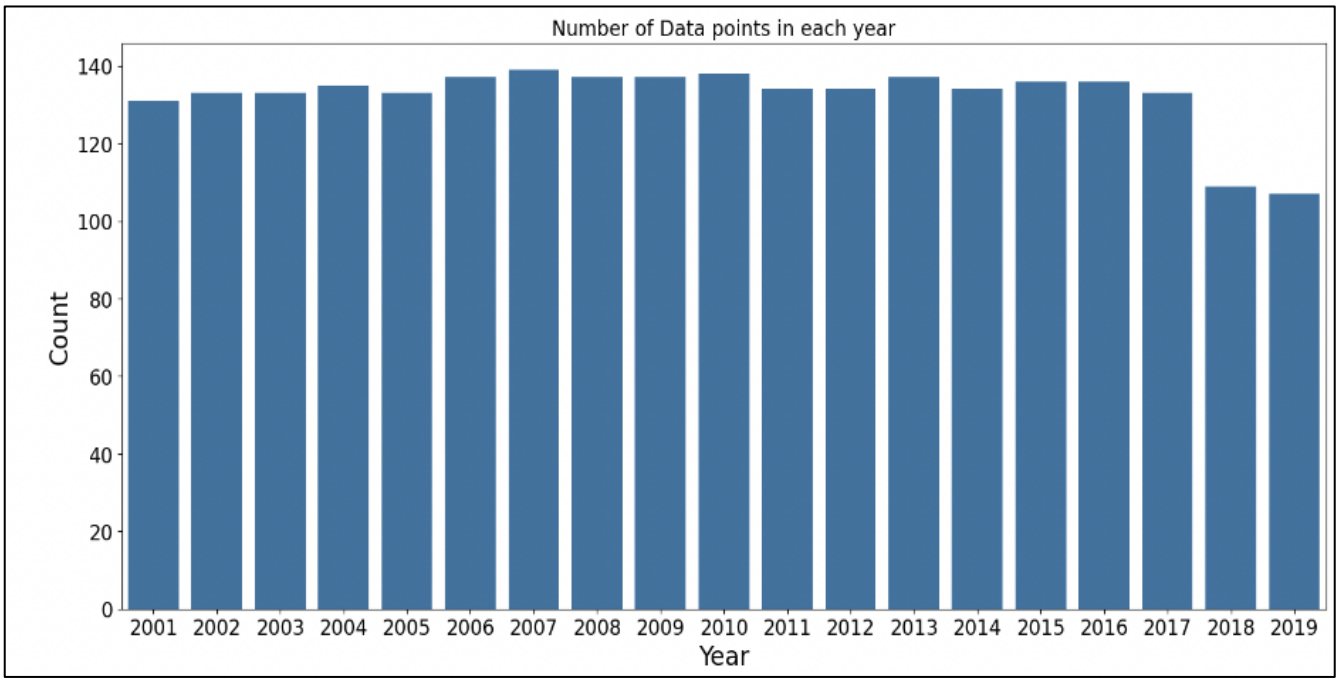
On the other hand, there are no duplicated data points in this dataset. Moreover, outlier handling has not been performed as the data follows a time series. Given the nature of time series data, variables tend to fluctuate over time, increasing the likelihood of normal values being mistakenly considered outliers. Additionally, this dataset includes 145 countries' data and each country has a different range of values so some countries' figures can be identified as outliers.

In summary, following data preprocessing, a total of 2,513 data points with 39 variables were remained and saved. The period ranges from 2001 to 2019.

3-2. Exploratory Data Analysis

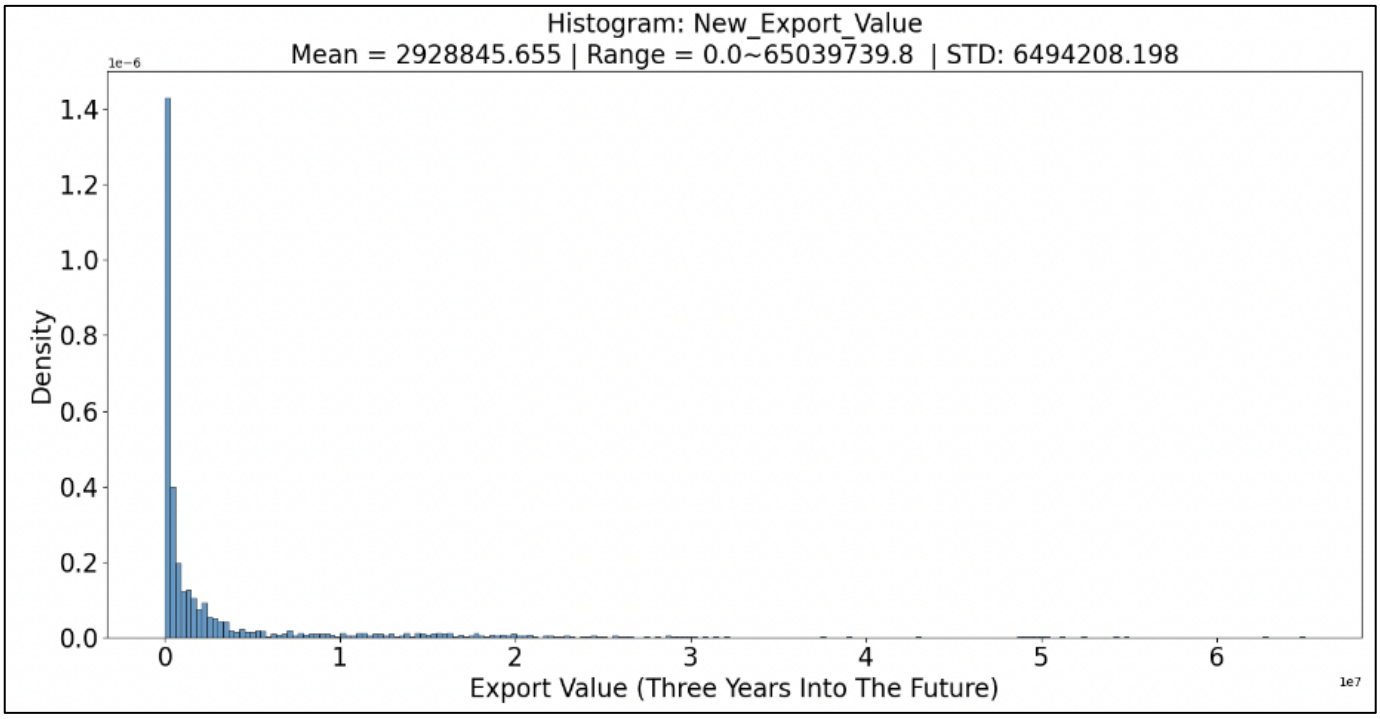
To begin with, the number of data points for each year was examined. As depicted in Figure 1, all years exhibit more than 100 data points. This suggests that all data points can be utilised for training and testing the MLP model.

Figure 1. Number of Data Points in Each Year



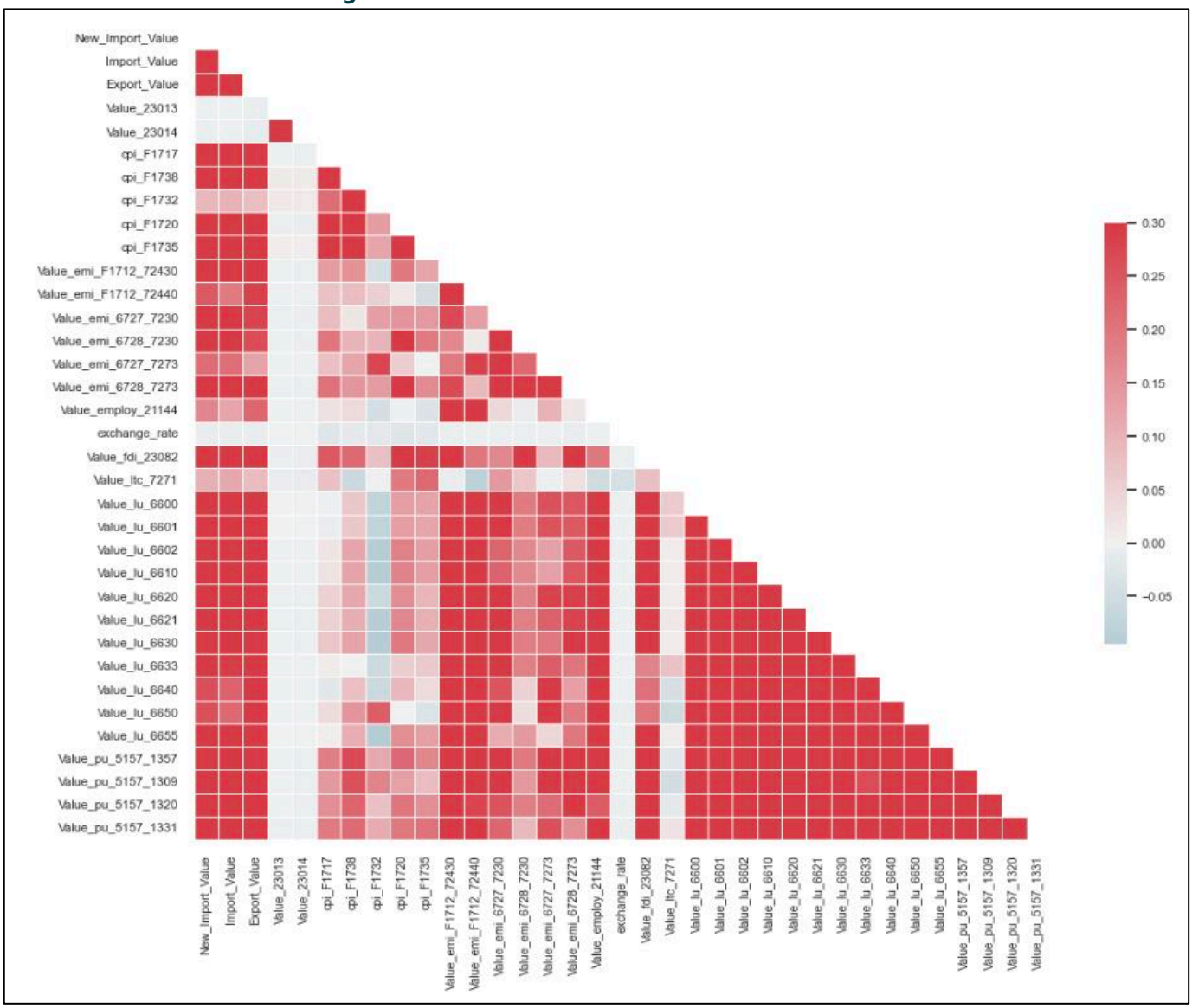
Moreover, the distribution of the dependent variable was examined. According to Figure 2, the 'New_Export_Value' variable ranges from 0 to 65,039,739.8 with a mean of 2,928,845.65 and a standard deviation of 6,494,208.198, a unit of 1000 US Dollars. This variable seems to have several outliers because each country has a different value. For example, developed countries such as the United States of America, and France tend to have high export values, average of 19,889,093.00 and 46,700,925.80 respectively. These large figures may appear as outliers.

Figure 2. Distribution of Dependent Variable



Afterwards, the data team investigated whether this dataset follows a statistical Gaussian distribution. Consequently, the null and alternative hypotheses were formulated, with the null hypothesis positing normal distribution and the alternative hypothesis suggesting non-Gaussian distribution. D'Agostino's K-squared and Lilliefors tests at a 1% significance level were performed to evaluate these hypotheses. The results revealed that none of the features followed a Gaussian distribution, as evidenced by the rejection of the null hypothesis for each feature. Furthermore, we visualised the distribution of each variable using histograms and an Empirical Cumulative Distribution Function (ECDF) plot. The ECDF plot's resemblance to the data distribution indicates whether the data follows a Gaussian distribution (Dekking and AI, 2005). The ECDF curve did not align closely with any specific data distribution. Appendix 2 represents the ECDF plots, and Appendix 3 describes the result of normality tests for all variables.

Figure 3. Correlation Coefficients Matrix



Furthermore, a correlation matrix has been drawn to identify relationships among the features. High positive or negative correlations between two features can adversely affect the results and outputs of data analysis and mathematical models due to multicollinearity (Gujarati and Porter, 2009). According to Figure 3, a lot of variables seem to be correlated with each other. This outcome may occur because some of the indicators and indices included in the variables tend to evaluate similar characteristics, potentially leading to a strong correlation value. To minimise the correlation between each variable, 12 variables with more than 80% correlations have been excluded. It is important to note that the 'New_Import_Value', 'Export_Value', and 'Import_Value' features were retained despite their significant correlation. This decision stemmed from the fact that 'New_Import_Value' is derived from 'Import_Value', and typically, the import and export values of a given product are intertwined.

- Correlated Variables: 'Value_23014','Value_emi_F1712_72430','Value_employ_21144','Value_emi_6728_7230','Value_lu_6600', 'Value_lu_6601','Value_lu_6602','Value_lu_6610', 'Value_lu_6620', 'Value_lu_6621', 'Value_pu_5157_1309', 'Value_pu_5157_1320'

Last but not least, the 'Area' variable was applied to one-hot encoding methods. One-hot encoding involves creating multiple binary variables that represent the presence or absence of different categories within a categorical variable. This process allows categorical data to be represented in a format suitable for machine learning algorithms. Thus, this model can learn the region's differences since this dataset now includes independent geographical variables, a total of 145 new binary variables. Following this, the 'Area' variable was deleted from the dataset. Additionally, the 'Year' variable is not used as an independent variable; therefore, it is not considered as such.

- Final Independent Variables: 'Value_23013', 'cpi_F1717', 'cpi_F1738', 'cpi_F1732', 'cpi_F1720', 'cpi_F1735', 'Value_emi_F1712_72440', 'Value_emi_6727_7230', 'Value_emi_6727_7273', 'Value_emi_6728_7273', 'exchange_rate', 'New_Import_Value', 'Value_fdi_23082', 'Value_ltc_7271', 'Value_lu_6630', 'Value_lu_6633', 'Value_lu_6640', 'Value_lu_6650', 'Value_lu_6655', 'Value_pu_5157_1357', 'Value_pu_5157_1331', 'Import_Value', and 'Export_Value' with 145 countries binary variables.

Nonetheless, this dataset consists of 168 independent variables and 1 dependent variable due to the one-hot encoding. Thus, it is recommended to conduct dimensionality reduction in order to reduce the complexity of data and computational cost.

3-3. Data Scaling and Dimensionality Reduction

In this dataset, each variable has a different range of values thus it needs to scale all variables. This practice proves beneficial as it mitigates the influence of variables with larger magnitudes, optimises the efficiency of model-based optimisation algorithms, and enhances the interpretability of the model's coefficients, among other advantages. For data scaling, Min-Max Scaler and Standard Scaler are often chosen to conduct it. The Min-Max Scaler scales data to a range between minimum and maximum values (Md. Johirul Islam et al., 2022).

Equation 3. Min-Max Scaler

$$\boldsymbol{Min - Max\ Normalisation} = \boldsymbol{x'} = \frac{\boldsymbol{x} - \boldsymbol{min(x)}}{\boldsymbol{max(x)} - \boldsymbol{min(x)}}$$

Where x represents the values of the variable to be normalised

Meanwhile, the Standard Scaler, using Z-score normalisation, transforms data to have a mean of 0 and a standard deviation of 1 (Fei et al., 2021). In this case, the Standard Scaler has been selected because the Min-Max Scaler is sensitive to outliers. The Standard Scaler was deemed more suitable since the data is in a time series format and outlier removal was not performed.

Equation 4. Standard Scaler

$$\boldsymbol{Standardisation} = \boldsymbol{x'} = \frac{\boldsymbol{x} - \boldsymbol{\bar{x}}}{\boldsymbol{\sigma}}$$

Where x represents the values of the variable to be standardised,

σ denotes the standard deviation of x and $\bar{x}$ means the mean of the variable x

Once the variables were scaled, a dimensionality reduction was undertaken using PCA. This technique simplifies data complexity while maintaining key characteristics by identifying the most influential patterns through a combination of variables that explain the highest variability (Jolliffe and Cadima, 2016). If these components represent around 70-90% of the population, it is considered suitable for machine learning. 135 components explain around 93% of the original data thus it can be concluded that proceeding with these components for further analysis is reasonable.

Consequently, after scaling and reducing dimensions, the dataset has been transformed from 168 features to 135 components. While 135 dimensions remain relatively high, this reduction from 168 features is expected to expedite computations and mitigate model complexity.

3-4. Dataset Split

The dataset has been randomly divided into a training dataset and a test dataset, with 80% allocated for training and 20% for testing. However, a validation dataset was not extracted from the training dataset because the data team implemented the cross-validation. Overall, the training and test datasets contain 2,010 and 503 data points respectively.

4. Data Modelling

4-1. Model Construction

A multilayer perceptron (MLP) neural network model has been selected to construct a forecasting model for the export value of crop products in Ethiopia and other regions. The MLP architecture consists of fully connected neurons with a nonlinear kind of activation function, arranged in at least three layers (Sakar et al., 2018). This carries out regression by training on labelled data, updating parameters through methods such as gradient descent, and minimising a loss function like mean squared error.

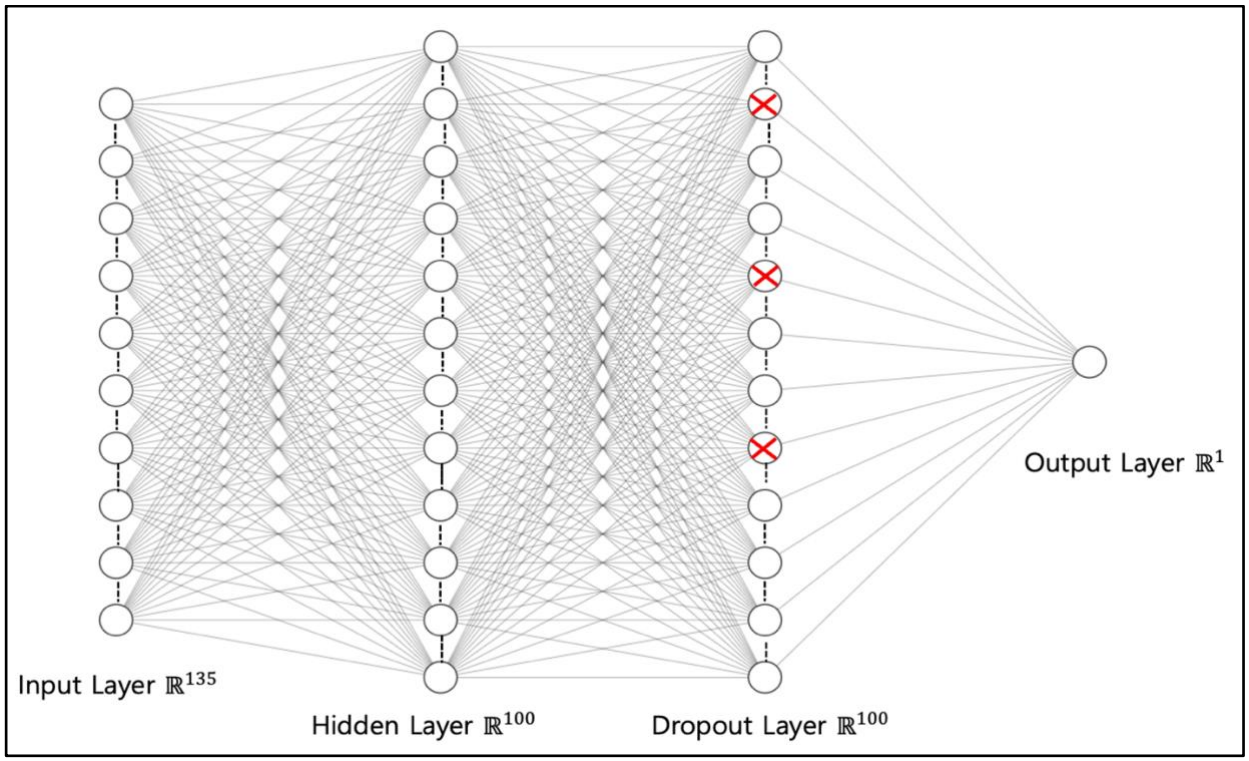
Table 2. Possible Hyperparameters

Hyperparameters	Value
Hidden Layer	10 | 30 | 50 | **100**
Learning Rate	0.001 | 0.01 | **0.1**
Optimizer	**Adam**

To determine the best combination of hyperparameters for the MLP model, K-fold cross-validation has been employed. This technique involves splitting the dataset into K subsets and iteratively training the model on different combinations of these subsets while evaluating its performance on the remaining data (Soper, 2021). By systematically rotating through these subsets, K-fold cross-validation provides a more robust estimate of the model's performance compared to a single validation set. It does not require a separate validation dataset and assists prevent overfitting.

Following the K-fold cross-validation with 5 folds and 500 epochs, the best combination identified consisted of 100 hidden layers, a learning rate of 0.1 and the Adam optimiser. Based on these hyperparameters, an MLP model was constructed as follows:

Figure 4. Structure of MLP Model



The forecasting model is made of an input layer, 100 hidden layers, a dropout layer, and an output layer. The input layer takes 135-dimensional data as input and fully connects the 100 hidden layers. The Rectified Linear Unit (ReLU) activation function in the hidden layers introduces nonlinearity by converting negative values as zeros.

Equation 5. ReLU Function

$$\boldsymbol{ReLu} = \boldsymbol{f(x)} = \boldsymbol{max(0, x)}$$

Where x represents the input data

In addition, the hidden layers apply L2 regularisation, restricting the weights to prevent the model from memorising the training data too closely, assisting it to avoid overfitting (Xu et al., 2017).

Equation 6. L2 Regularisation

$$\boldsymbol{L_2\ regularisation} = \|\boldsymbol{\omega}\|_2^2 = \boldsymbol{\omega_1^2 + \omega_2^2 + \cdots + \omega_n^2}$$

Where ω denotes the feature weights to be regularised

To further preclude overfitting, the dropout layer arbitrarily selects 30% of neurons to eliminate during each training iteration. Since this model handles a regression case, the output layer does not have an activation function.

Overall, several methods were opted to prevent the overfitting in this model:

① K-Fold Cross-Validation: To find the best set of hyperparameters.

② L2 regularisation: Constraining the weights to prevent the model from overfitting to the training data too closely. This model has a regularisation strength of 0.01.

③ Dropout Layer: Randomly deactivating a certain portion of neurons to prevent overfitting and enhance generalization. In this model, the dropout layer drops 30% of neurons.

Moreover, mean squared error (MSE) was chosen as a loss function to assess the model's performance and implement backpropagation.

Equation 7. Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

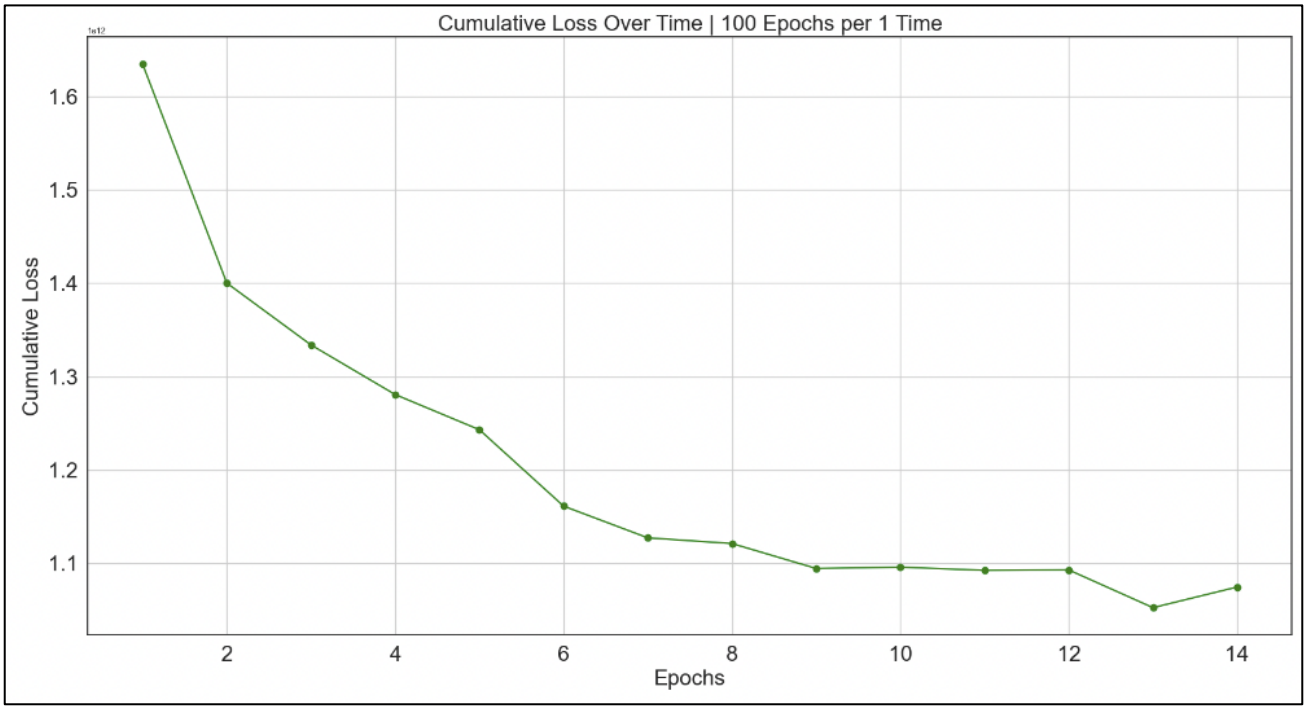
Where Y represents the values of the dependent variable,

$\hat{Y}$ denotes the predicted values of the dependent variable

4-2. Model Assessment

The MLP model underwent a total of 1,400 epochs. This model was trained through the training dataset including 2,010 data points. Afterwards, the test dataset with 503 data points evaluated the model. The total epochs were determined based on the test loss observed in Figure 5. At the end of the training, the test loss was quite stabilised, indicating a convergence of the model's performance. Therefore, the training process was halted.

Figure 5. Cumulative Test Loss



The final model's MSE, root mean squared error (RMSE), and mean absolute error (MAE) have been calculated to assess this model. The formulas for RMSE and MAE are as follows (Willmott and Matsuura, 2005):

Equation 8. RMSE and MAE

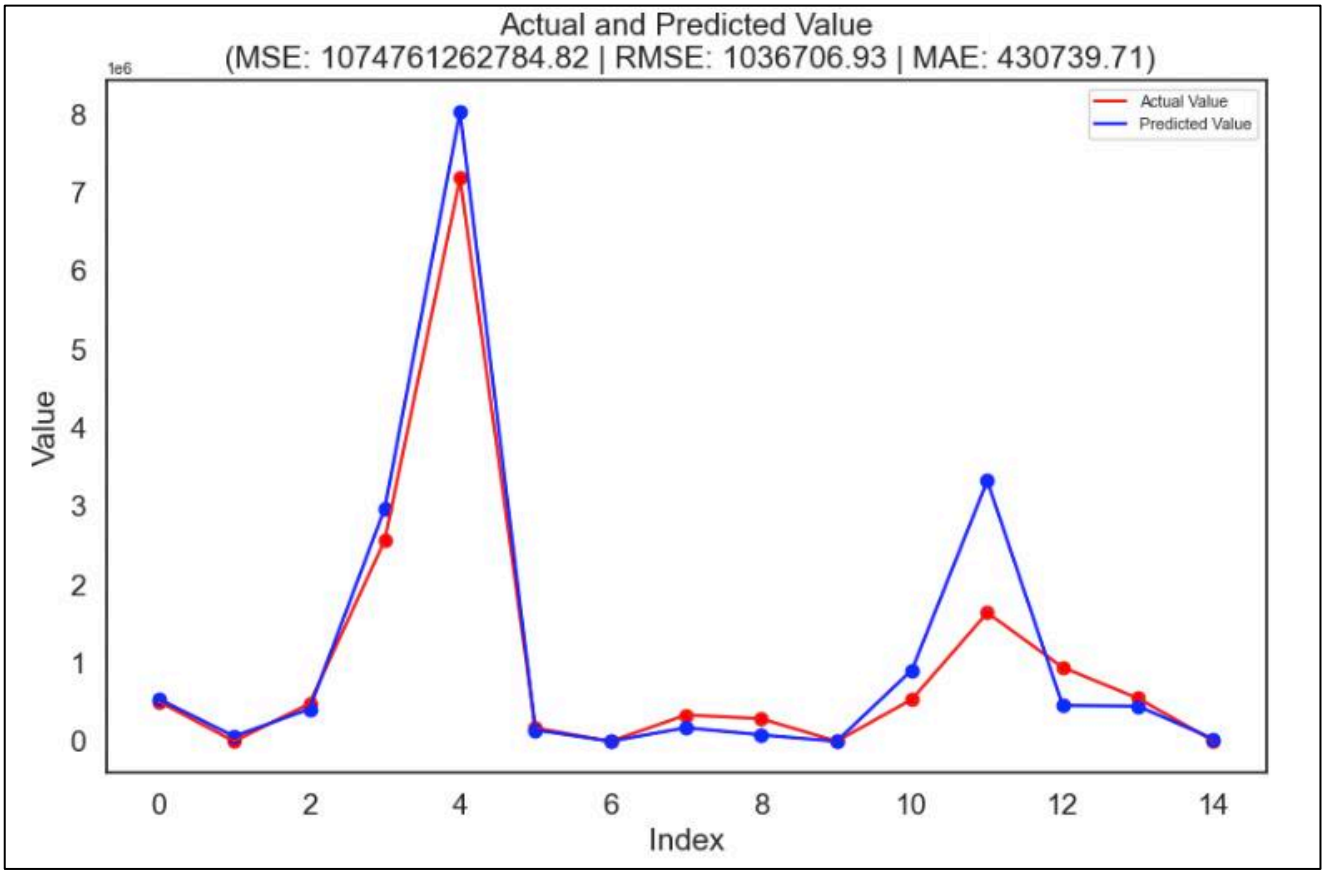
$$RMSE = \sqrt{MSE} \quad | \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y represents the values of the dependent variable,

$\hat{y}$ denotes the predicted values of the dependent variable

Consequently, the model yielded an MSE of 1,074,761,262,784.82, an RMSE of 1,036,706.93, and an MAE of 430,739.71. Since the data range from 0 to 65,039,739.8 with a mean of 2,928,845.65, it is possible to have a relatively high MSE even with small estimation errors. In this case, although the MSE value is substantial, the RMSE of 1,036,706.93 and the MAE of 430,739.71 suggest that the errors are not significantly high compared to the scale of the data.

Figure 6. Result of Model Assessment



Besides, upon analysing the model assessment in Figure 6 that randomly extracted 15 data points from the test dataset, it becomes evident that more than half of the predicted values are nearly

aligned with actual values. This means that the model is effectively performing in forecasting the export value of crop products across most regions.

Table 3. Forecasting Result

Country	Year	Actual Value (1000$)	Prediction (1000$)
Ethiopia	2022	568,993.96	613,309.75
Republic of Korea	2022	2,408,408.80	2,329,309.75

Finally, this model was assessed through the independent variables from Ethiopia and the Republic of Korea in 2019. Given the aim for broad applicability, the data team used features from two highly contrasting countries Ethiopia and the Republic of Korea in terms of economy and geography. As a result, this model forecasted that the crop products' export value in Ethiopia and Korea in 2022 will be $613,309,750 and $2,329,309,750 respectively. Comparing these forecasts with the actual values of $568,993,960 and $2,408,408,800 for Ethiopia and Korea in 2022, it can be concluded that the model demonstrated a high level of accuracy in predicting crop products' export values considering together with the aforementioned MSE, RMSE, and MAE values. This implies that the model is robust and capable of providing precise forecasts for various regions.

5. Conclusion

This project built the MLP model that forecasts the export values of crop products three years later in Ethiopia and other countries. After data pre-processing, the data was scaled through Standard Scaler and dimensionality reduction was performed by converting 168 features into 135 components using PCA. Furthermore, K-fold cross-validation was conducted to find the best set of hyperparameters. Following this, the model was developed based on the hyperparameters.

The model was trained 1,400 times and assessed by MSE, RMSE, and MAE. Even if MSE (1,074,761,262,784.82) is high, RMSE and MAE are not significantly large under the scale of the data, each holds 1,036,706.93 and 430,739.71. Using data from Ethiopia and the Republic of Korea, the model predicted crop product export values for 2022, with forecasts closely matching actual figures. This highlights the model's robustness and ability to provide useful forecasts across diverse regions. However, this model does not focus on one specific region so it might not be inaccurate in some regions.

In conclusion, the client is able to predict the export value of crop products in not only Ethiopia but also in other countries through this model. This allows them to formulate policies, strategies, and relevant programmes regarding the crop products business through this expected value.

6. References

- Dekking, M. and Al, E. (2005). A modern introduction to probability and statistics / understanding why and how. London: Springer. p. 219. ISBN 978-1-85233-896-1
- Gujarati, D.N. and Porter, D.C. (2009). "Multicollinearity: what happens if the regressors are correlated?". Basic Econometrics (4th ed.). McGraw–Hill. pp. 363. ISBN 9780073375779
- Md. Johirul Islam, Ahmad, S., Haque, F., Mamun, Arif, M. and Md. Rezaul Islam (2022). Application of Min-Max Normalization on Subject-Invariant EMG Pattern Recognition. IEEE Transactions on Instrumentation and Measurement, 71, pp.1–12. doi:https://doi.org/10.1109/tim.2022.3220286
- Fei, N., Gao, Y., Lu, Z. and Xiang, T. (2021). Z-Score Normalization, Hubness, and Few-Shot Learning. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). doi:https://doi.org/10.1109/iccv48922.2021.00021
- Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. doi:https://doi.org/10.1098/rsta.2015.0202
- Sakar, C.O., Polat, S.O., Katircioglu, M. and Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31(10), pp.6893–6908. doi:https://doi.org/10.1007/s00521-018-3523-0
- Soper, D.S. (2021). Greed Is Good: Rapid Hyperparameter Optimization and Model Selection Using Greedy k-Fold Cross Validation. Electronics, 10(16), 1973. doi:https://doi.org/10.3390/electronics10161973.
- Xu, Y., Zhong, Z., Yang, J., You, J. and Zhang, L. (2017). A New Discriminative Sparse Representation Method for Robust Face Recognition via L2 Regularization. 28(10), pp.2233–2242. doi:https://doi.org/10.1109/tnnls.2016.2580572
- Willmott, C. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), pp.79–82. doi:https://doi.org/10.3354/cr030079

7. Appendix

7-1. List of CSV*

Source: ML Coursework Dataset

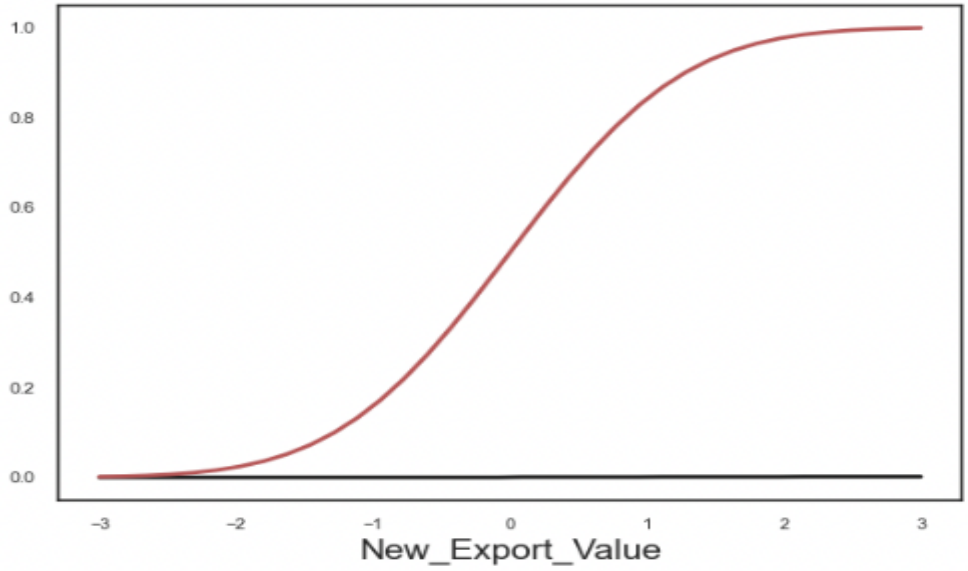
Data file	Summary of variables (variable units are indicated in the file, where applicable)
Consumer price indicators	• Consumer price, food index • Country • Food price inflation • Month • Year
Crops production indicators	• Country • Year • Yield for different crop products
Emissions	• Country • Crops CH4 emissions • Crops N2O emissions • Drained soil CO2 emissions • Drained soil N2O emissions • Year
Employment	• Country • Employment (male and female total) in agriculture, forestry, and fishing • Mean weekly hours worked per person (no distinction between male and female) in agriculture, forestry, and fishing • Year
Exchange rate	• Country • Currency • Local currency units per USD • Months • Year
Fertilizers use	• Agricultural use of different categories of fertilizers • Country • Year

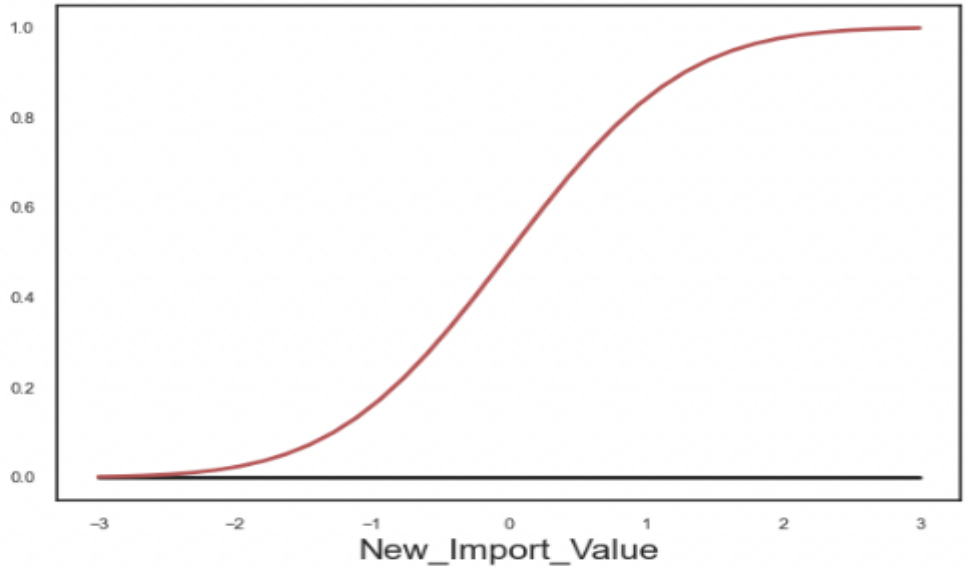
Food balances	• Country • Export quantity for different crop and livestock products • Food uses for different crop and livestock products • Import quantity for different crop and livestock products • Losses for crop and livestock products • Other uses for crop and livestock products • Year
Food security	• Cereal import dependency ratio • Country • Dietary energy supply adequacy • Per capita food production variability • Per capita food supply variability • Percentage of arable land equipped for irrigation • Political stability and absence of violence/terrorism index • Prevalence of anaemia in women of reproductive age • Prevalence of low birthweight • Protein energy supply • Value of food imports in total merchandise exports • Year
Food trade	• Country • Export value • Import value • Year
Foreign direct investment (FDI)	• Country • FDI inflows to agriculture, forestry, and fishing • FDI inflows to food, beverages, and tobacco • FDI outflows to agriculture, forestry, and fishing • FDI outflows to food, beverages, and tobacco • Total FDI inflows • Total FDI outflows • Year
Land temperature change	• Country

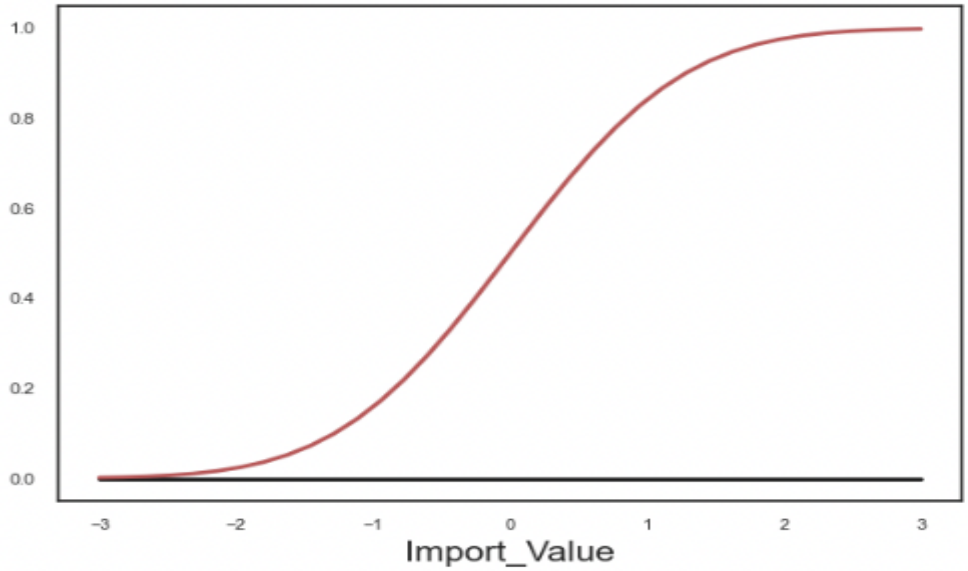
	• Months • Temperature change • Standard deviation • Year
Land use	• Area for different categories of land use • Country • Year
Pesticides use	• Agricultural use for each of fungicides (and bactericides), herbicides, insecticides, pesticides, rodenticides • Country • Use per area of cropland for each of fungicides (and bactericides), herbicides, insecticides, pesticides, rodenticides • Use per value of agricultural production for each of fungicides (and bactericides), herbicides, insecticides, pesticides, rodenticides • Year

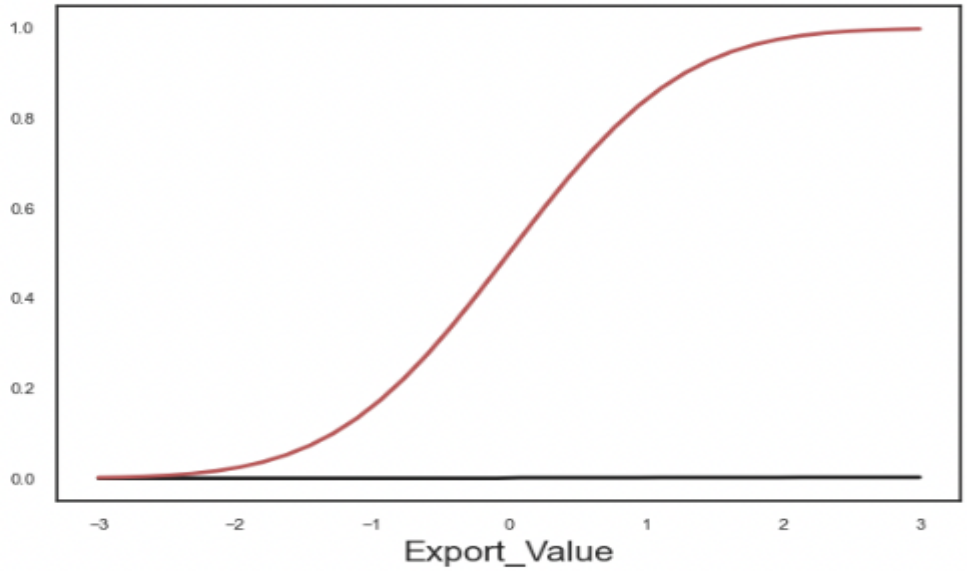
7-2. ECDF Plots

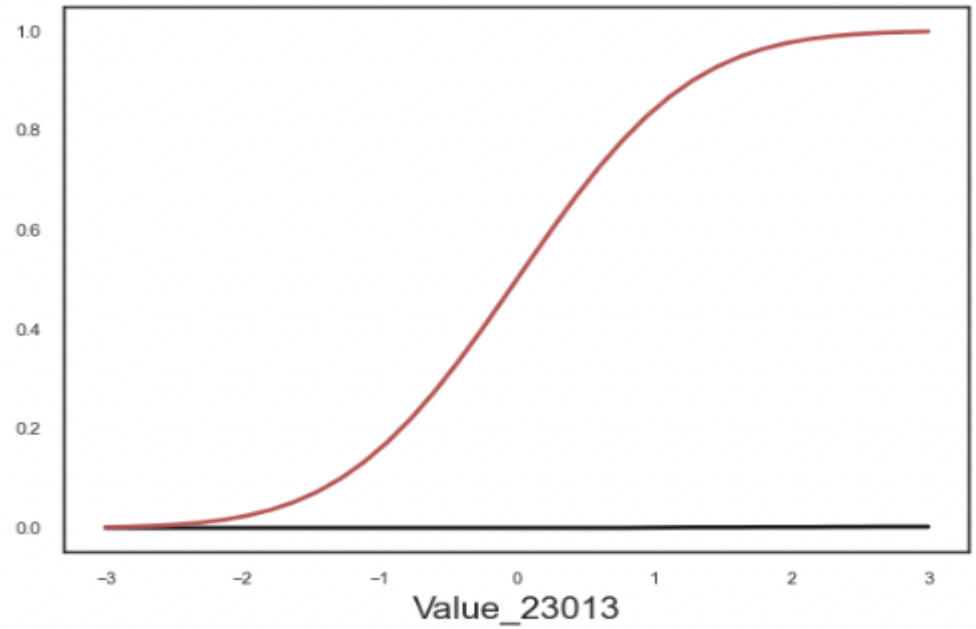
ECDF Plots

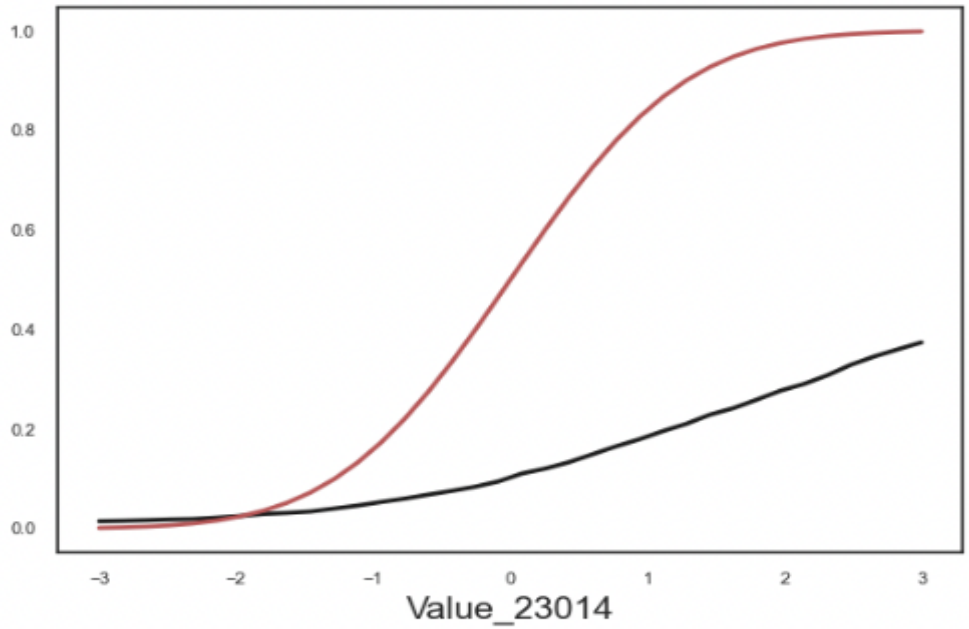


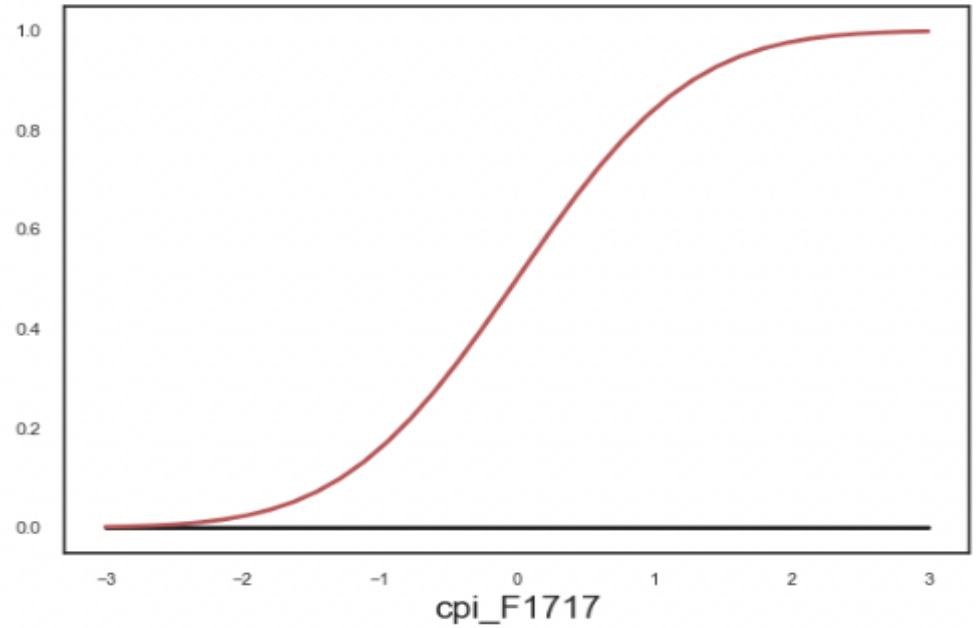


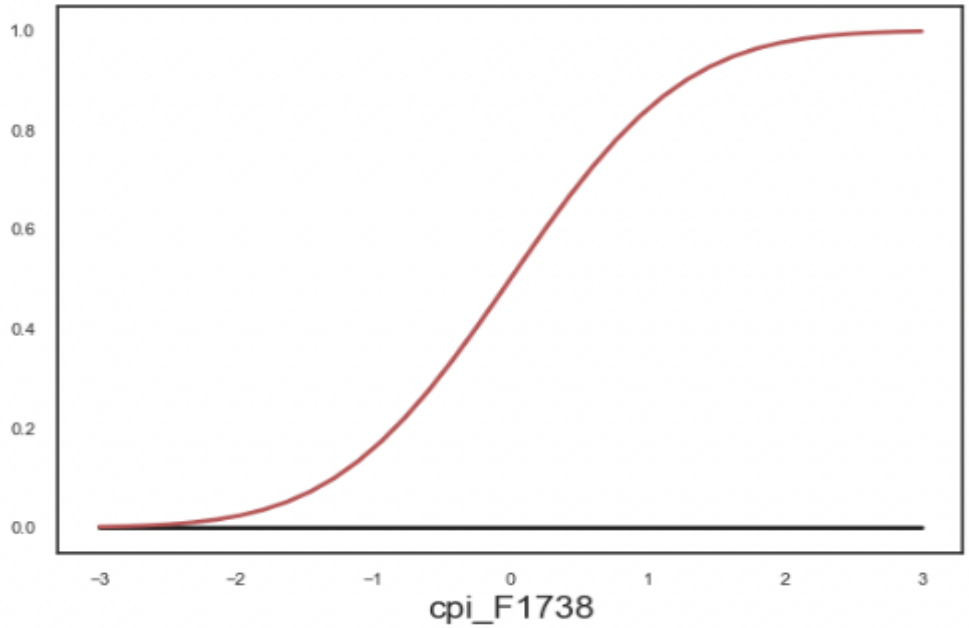


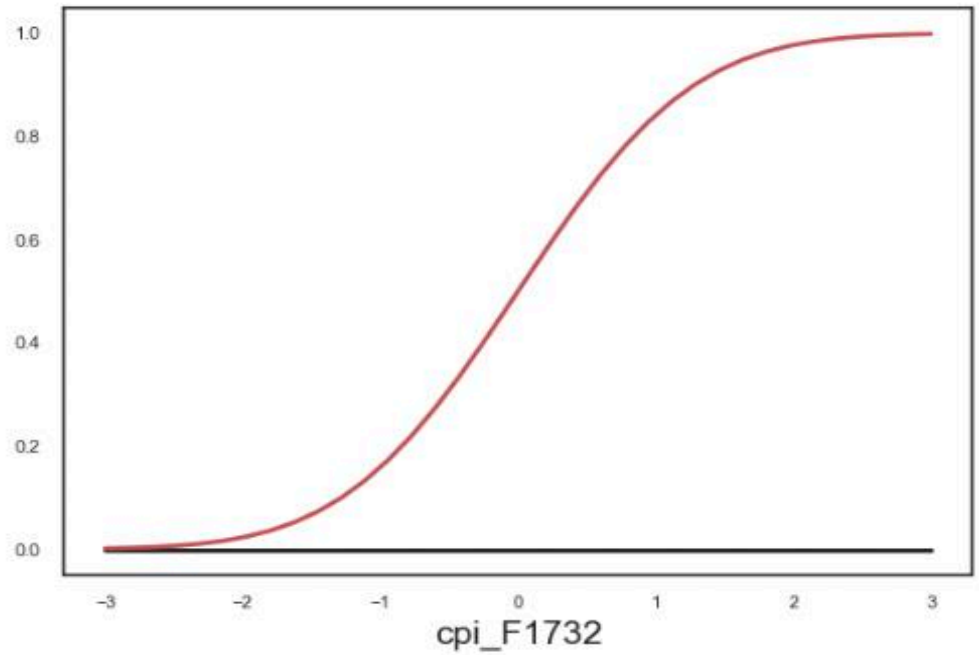


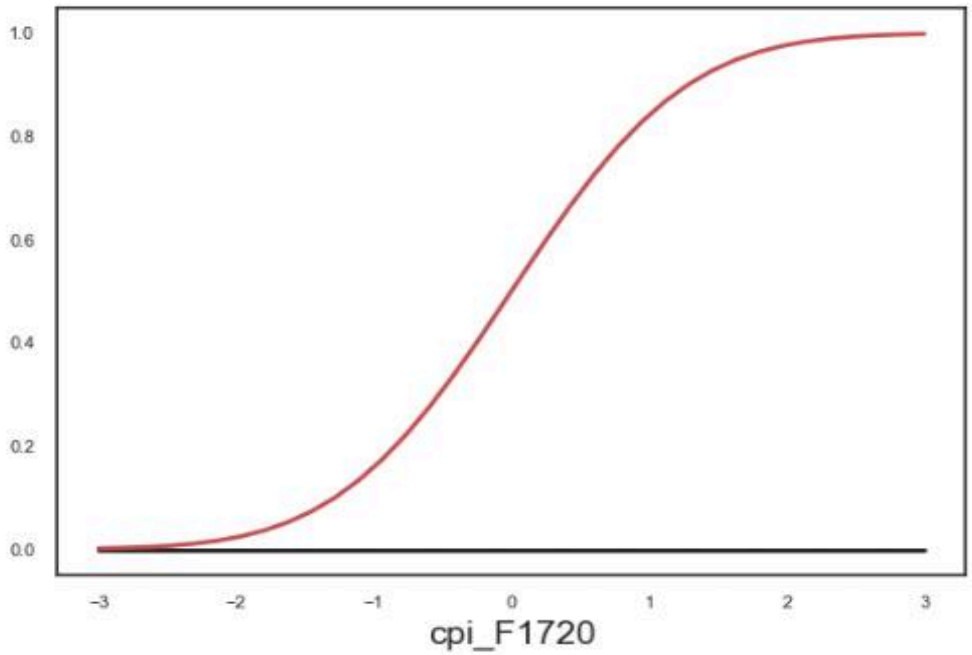

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_23013

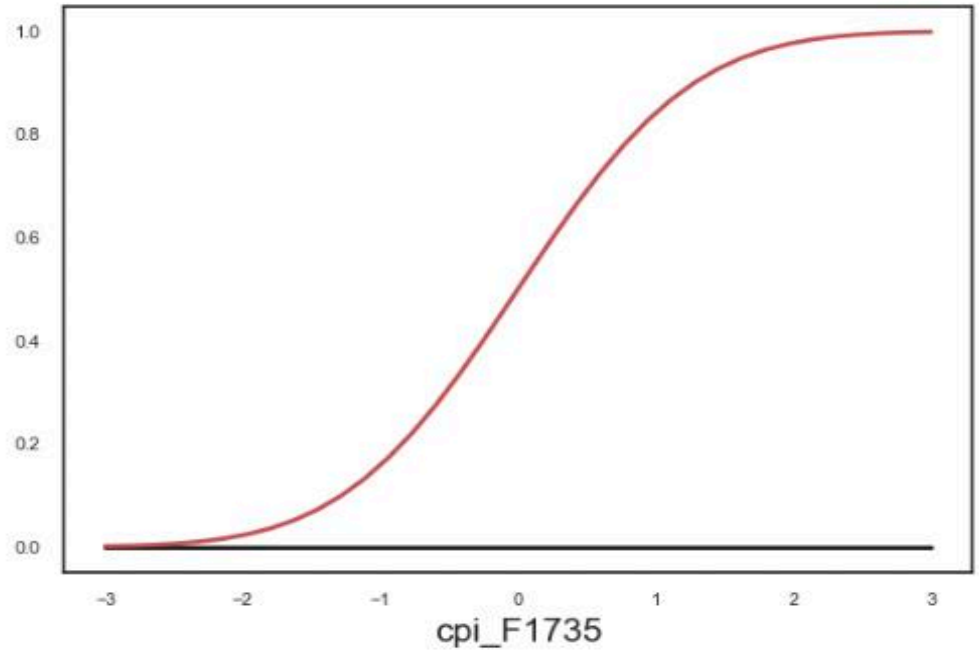

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_23014

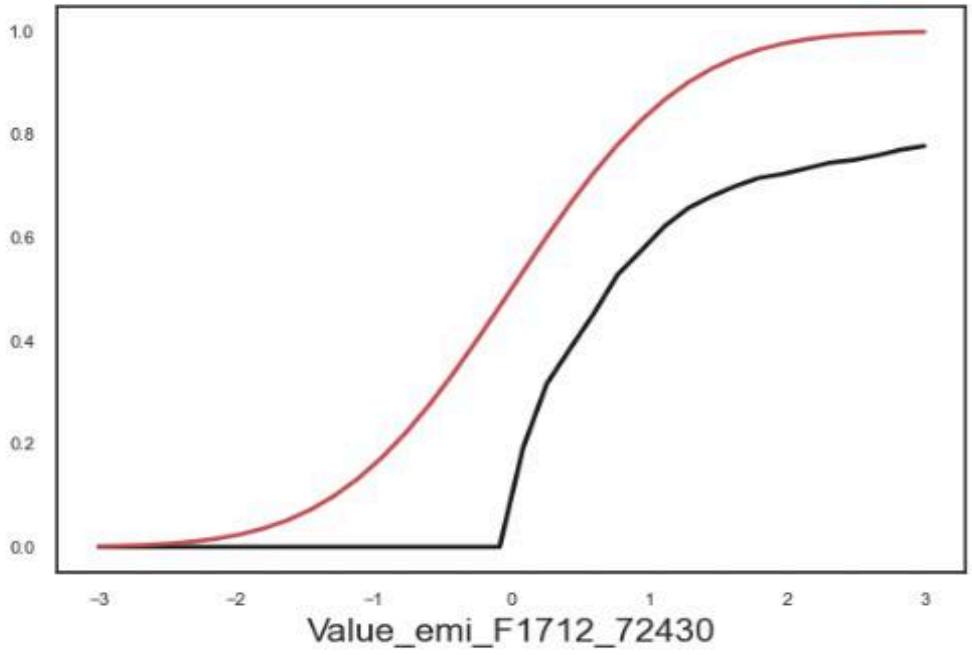

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
cpi_F1717

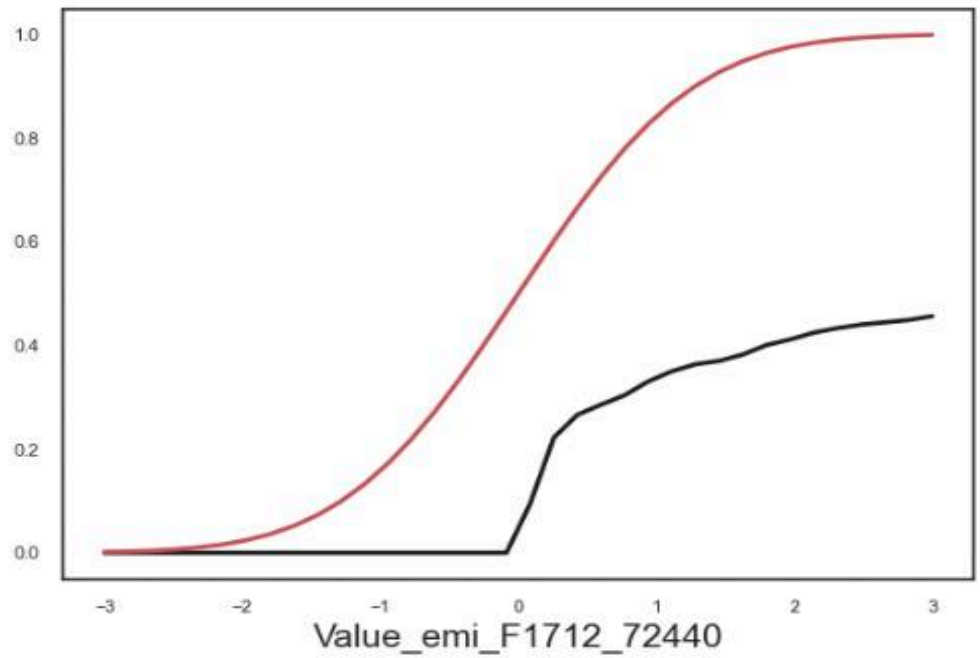

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
cpi_F1738

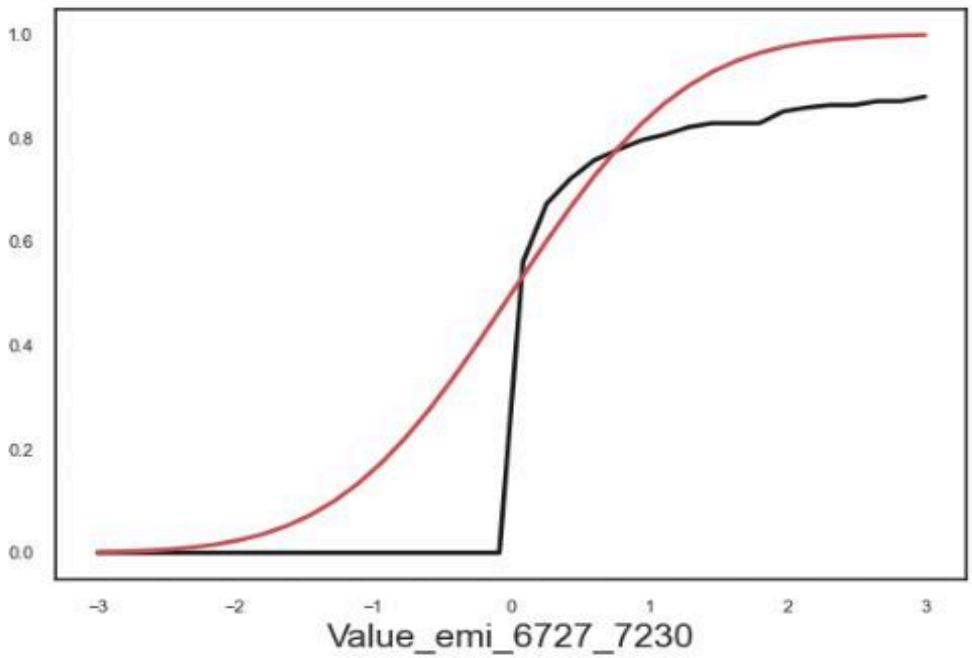

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
cpi_F1732

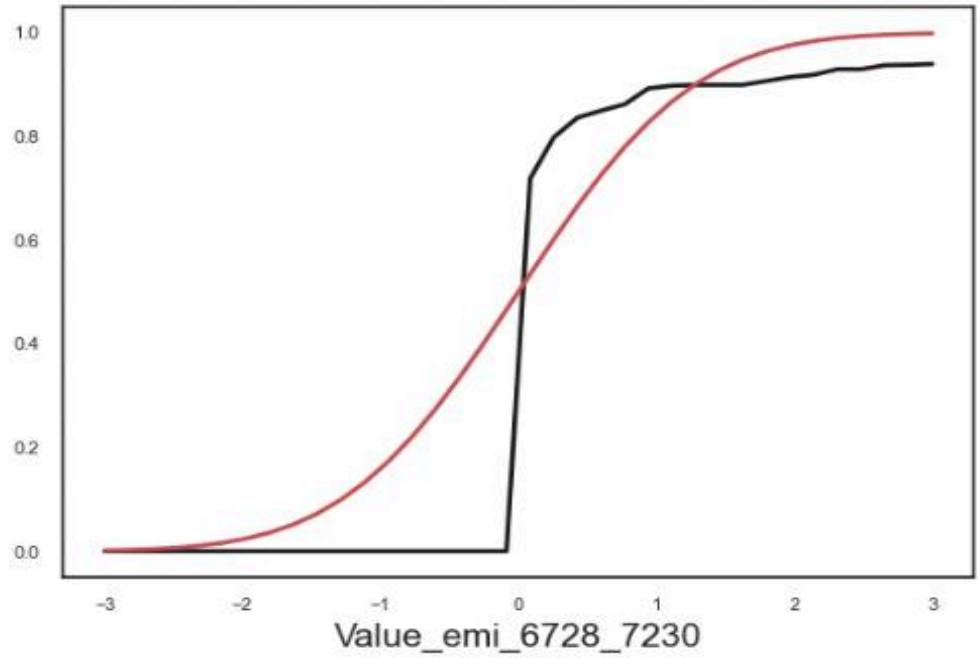

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
cpi_F1720

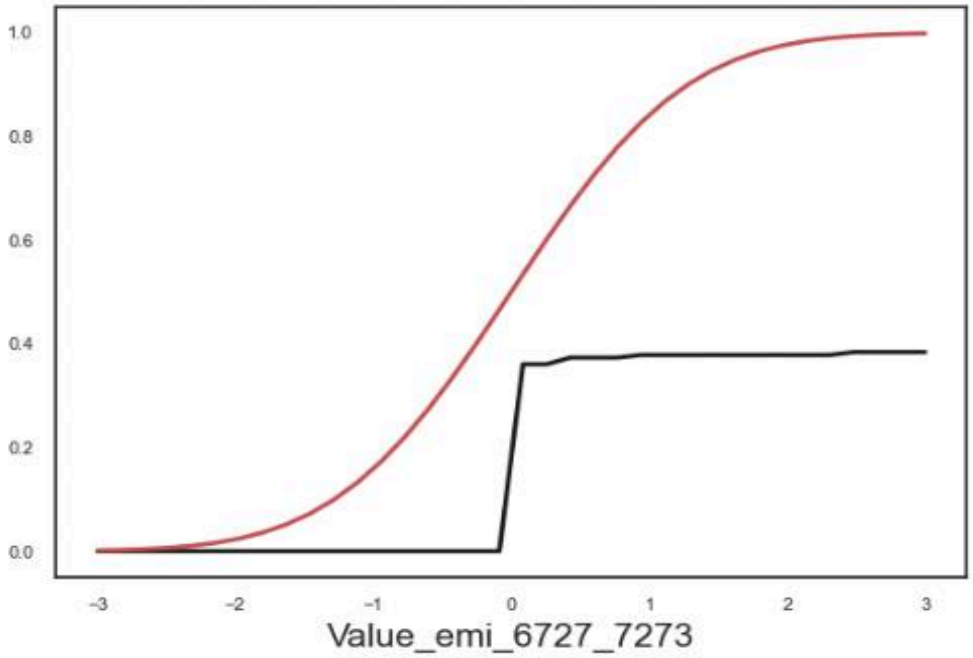

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
cpi_F1735

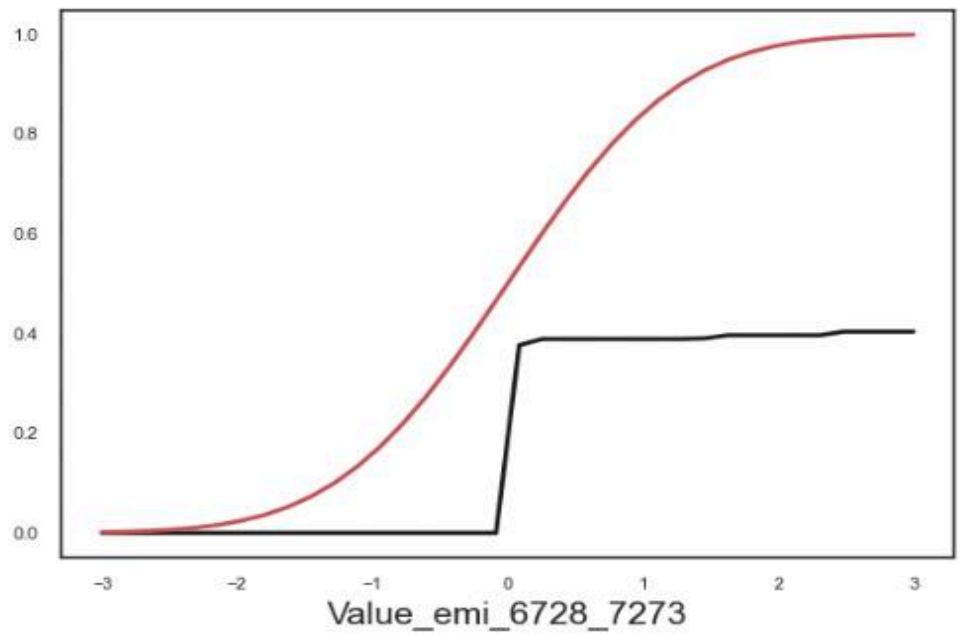

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_emi_F1712_72430

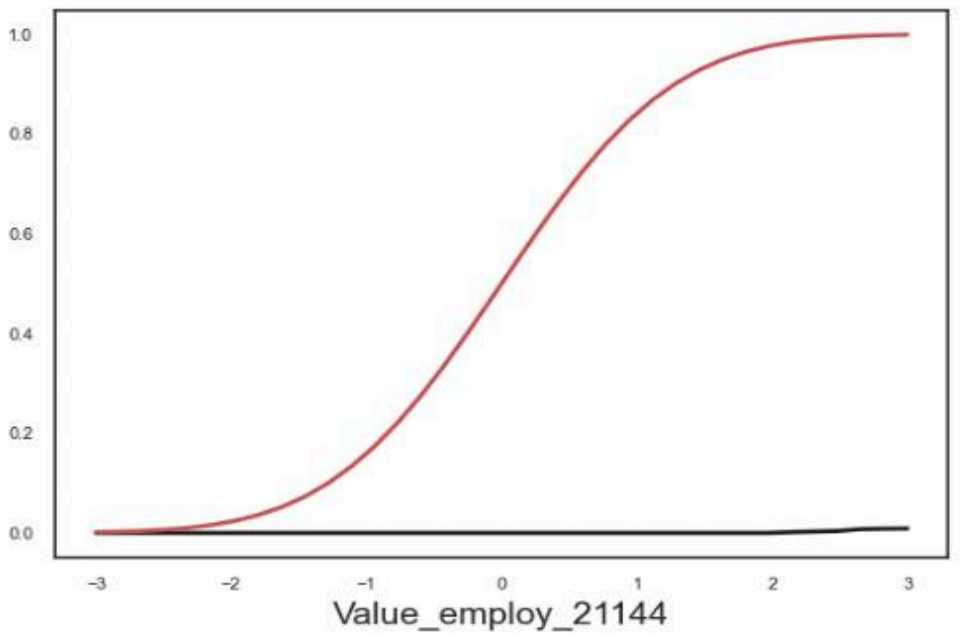

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_emi_F1712_72440

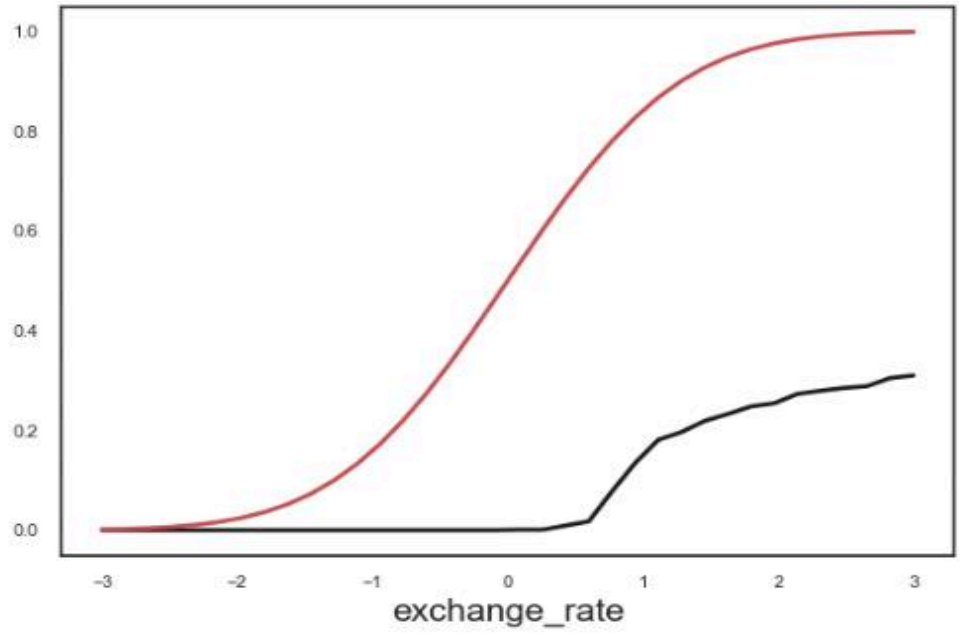

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_emi_6727_7230

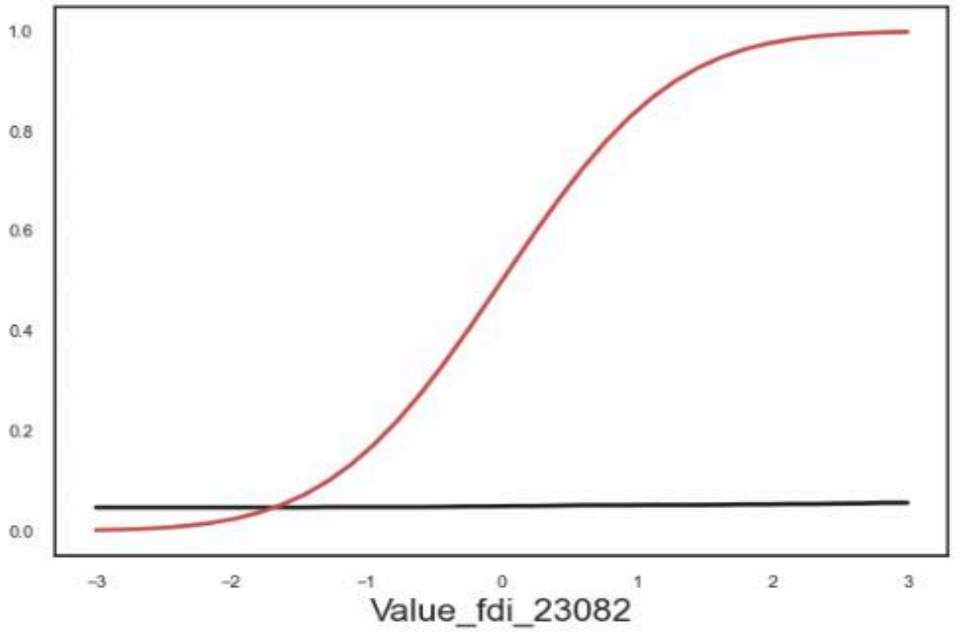

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_emi_6728_7230

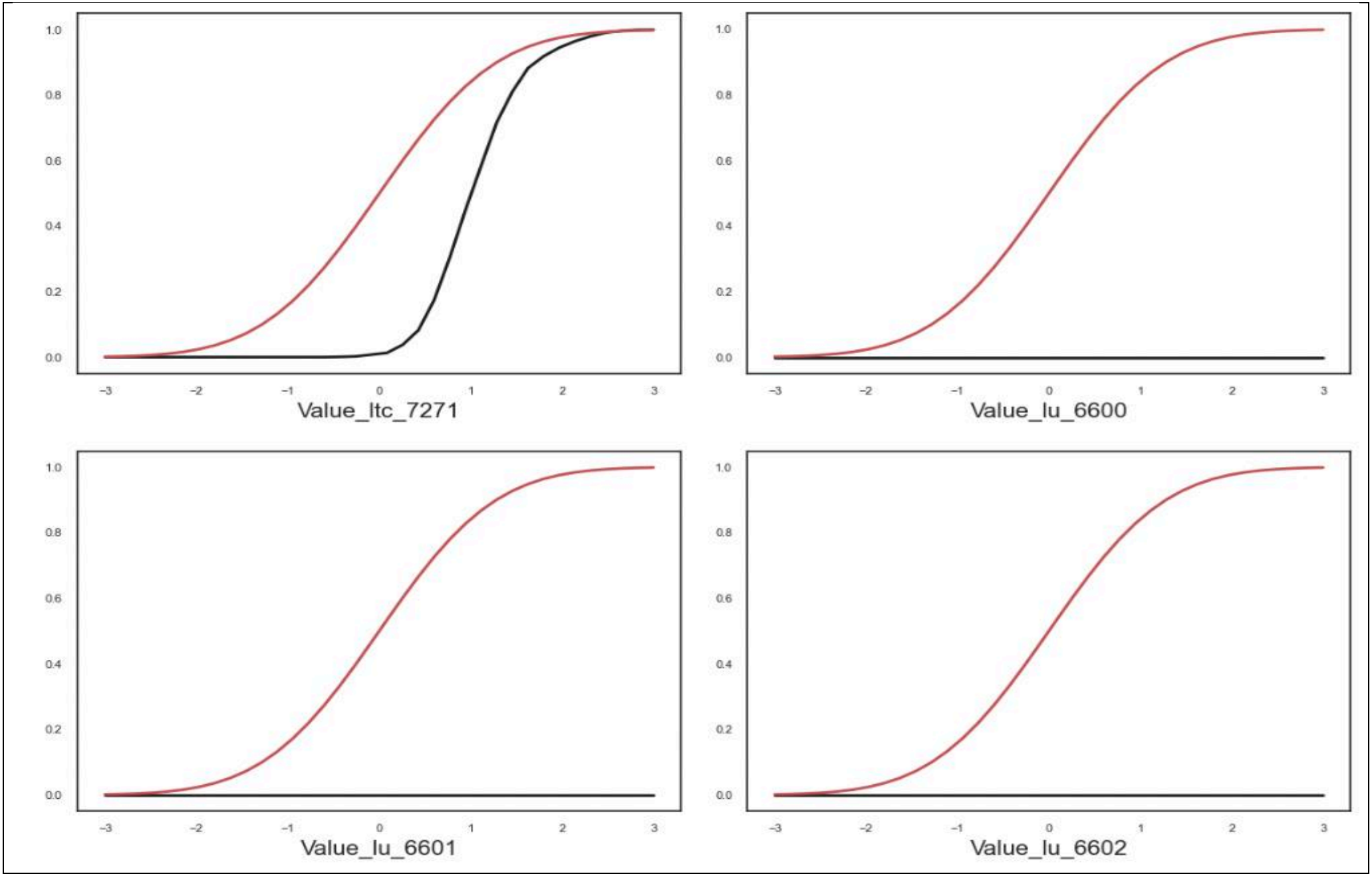

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_emi_6727_7273

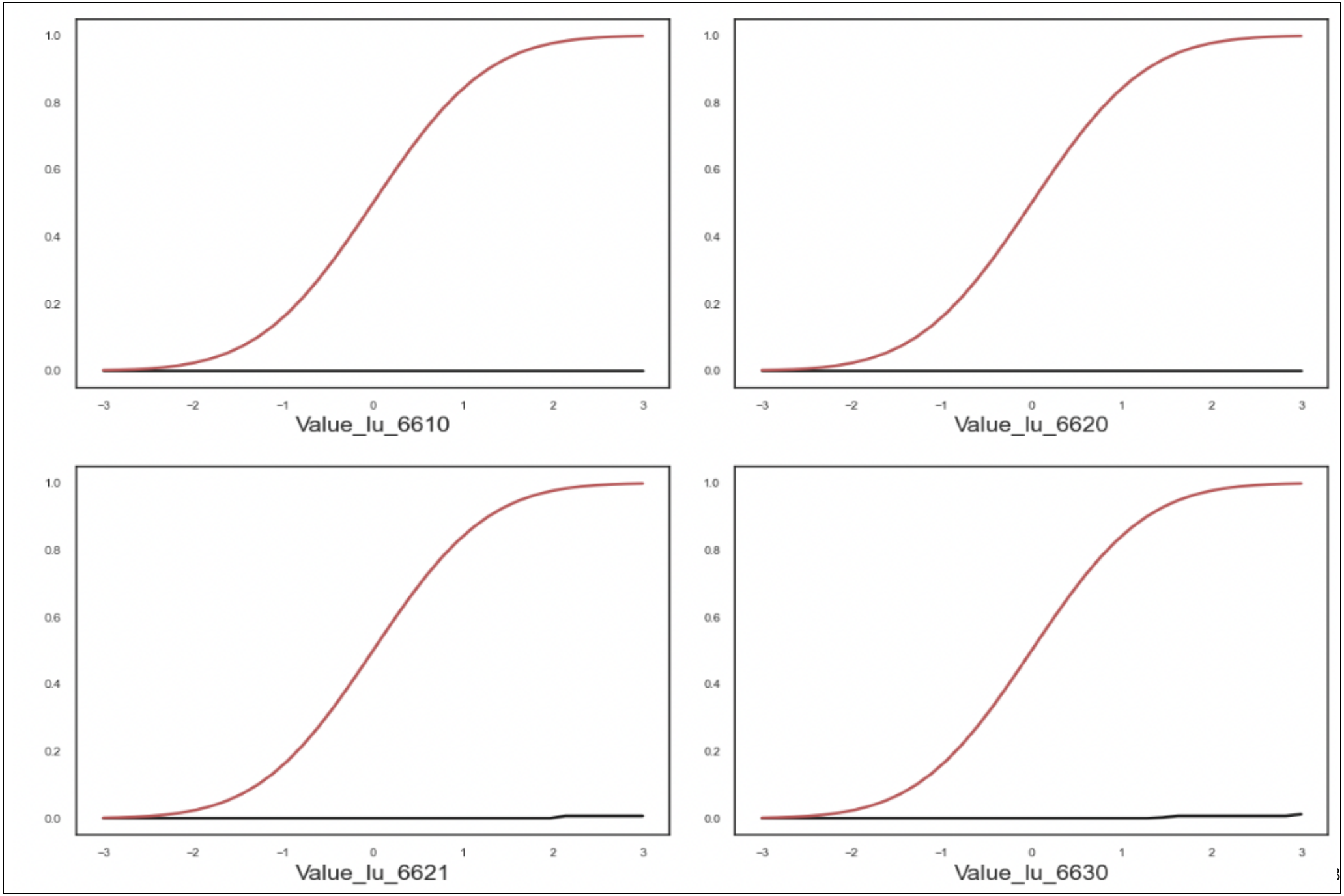

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_emi_6728_7273

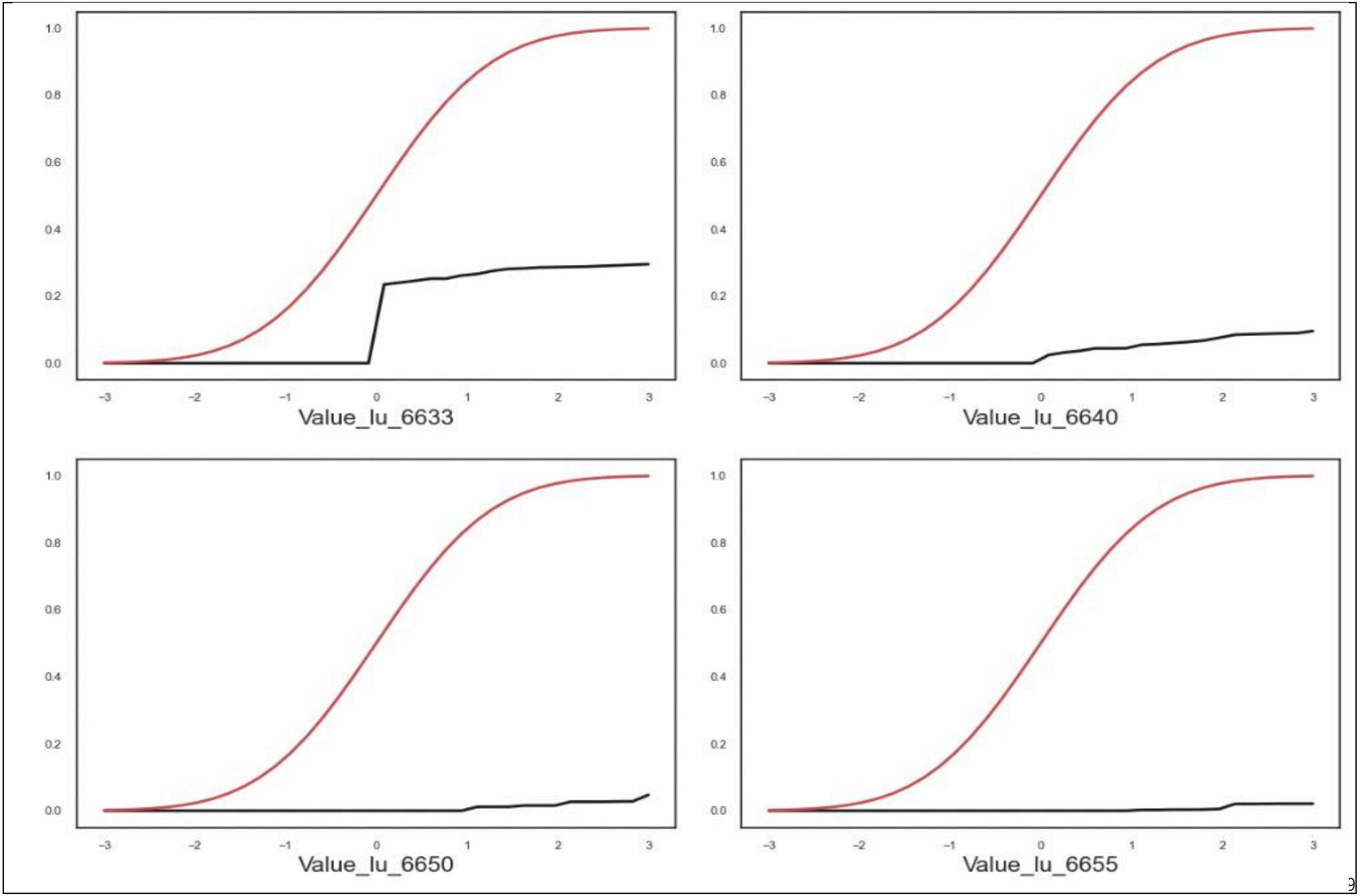

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_employ_21144

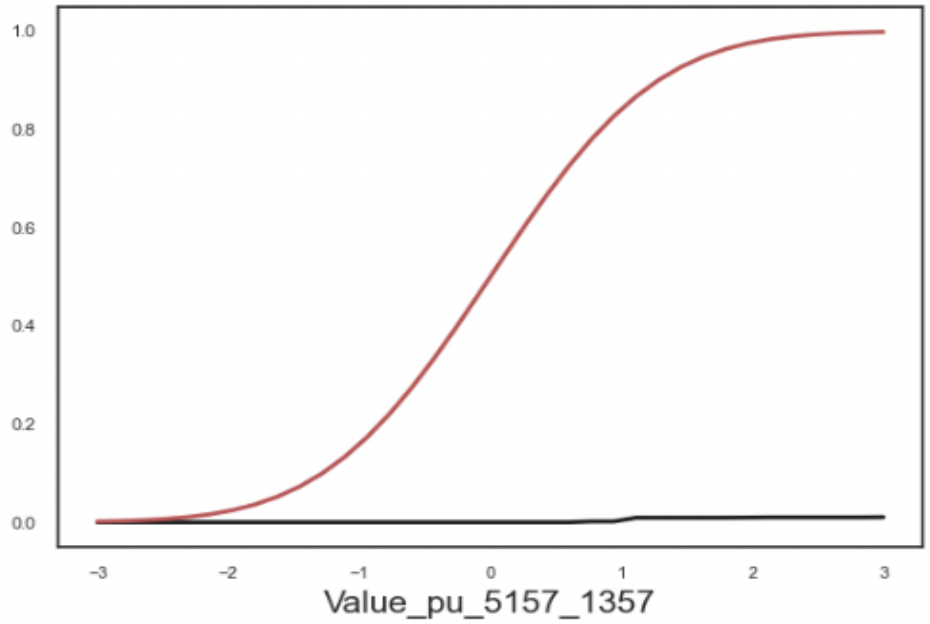

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
exchange_rate

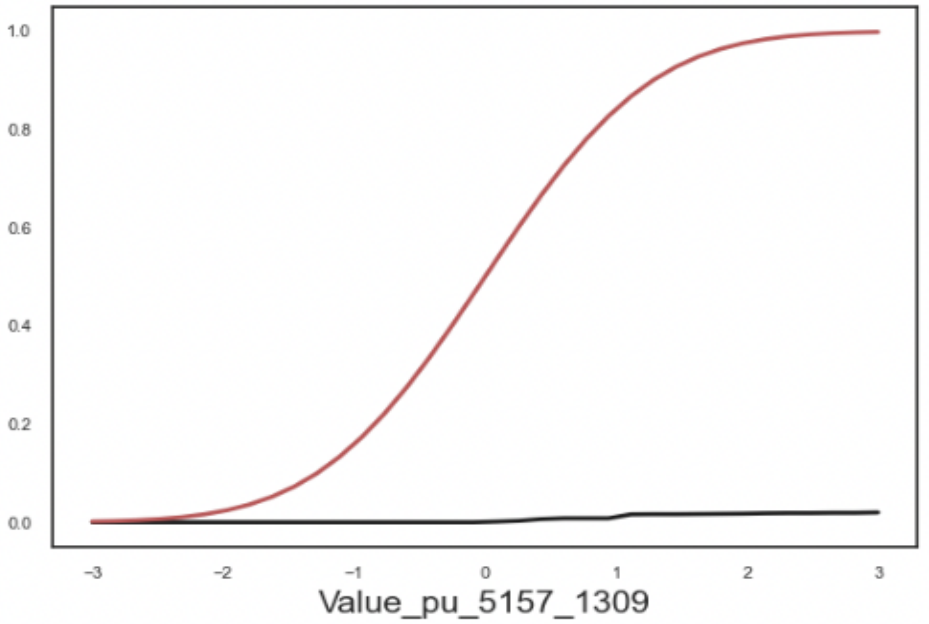

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_fdi_23082

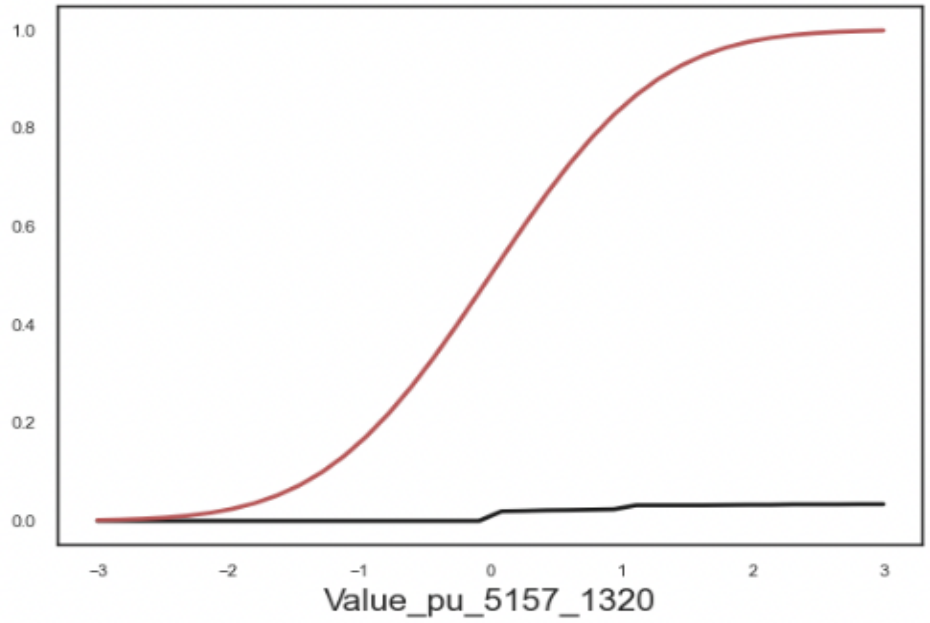

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_ltc_7271
Value_lu_6600
Value_lu_6601
Value_lu_6602

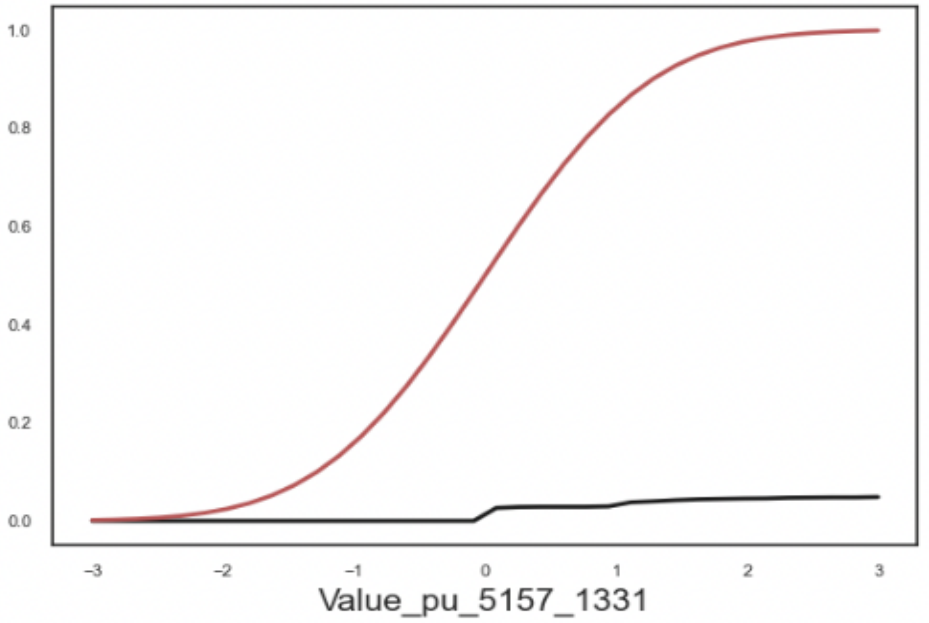

1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_lu_6610
Value_lu_6620
Value_lu_6621
Value_lu_6630


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_lu_6633
Value_lu_6640
Value_lu_6650
Value_lu_6655


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_pu_5157_1357


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_pu_5157_1309


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_pu_5157_1320


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Value_pu_5157_1331

7-3. Result of Normality Tests

Normality Test		
Null Hypothesis	This variable follows a statistical Gaussian distribution	
Alternative Hypothesis	This variable does not follow a statistical Gaussian distribution	
	p-value ≥ Significance Level (0.01)	**p-value < Significance Level (0.01)**
D'Agostino's Ksquared	None	'Value_23013', 'Value_23014', 'cpi_F1717', 'cpi_F1738', 'cpi_F1732', 'cpi_F1720', 'cpi_F1735', 'Value_emi_F1712_72430', 'Value_emi_F1712_72440','Value_emi_6727_7230', 'Value_emi_6728_7230','Value_emi_6727_7273', 'Value_emi_6728_7273', 'Value_employ_21144', 'exchange_rate', 'New_Import_Value', 'Value_fdi_23082', 'Value_Itc_7271', 'Value_lu_6600', 'Value_lu_6601', 'Value_lu_6602', 'Value_lu_6610', 'Value_lu_6620', 'Value_lu_6621', 'Value_lu_6630', 'Value_lu_6633', 'Value_lu_6640', 'Value_lu_6650', 'Value_lu_6655', 'Value_pu_5157_1357', 'Value_pu_5157_1309', 'Value_pu_5157_1320', 'Value_pu_5157_1331', 'New_Export_Value', 'Export_Value', 'Import_Value'
Lilleifors		