Multilayer Perceptron Model for Projecting Crop Product Export Value Forecasting

Introduction

In order to predict the export value of agricultural products for a specific geographic area three years into the future, this report details the development and evaluation of a Multilayer Perceptron (MLP) model. The goal is to forecast export values for the year 2987, using historical data available up to the year 2984. The report is structured into four main sections: Performance, Features & Labels, MLP Model, and Pre-processing.

The process begins with thorough data pre-processing, which includes handling missing values, standardizing the data, and splitting the dataset into training and testing sets. Following this, the MLP model is trained on the pre-processed data and evaluated to ensure it can predict future export values accurately and reliably. The comprehensive procedure aims to ensure the model's effectiveness in making precise forecasts for future agricultural export values..

Performance

The performance of my regression model was evaluated using the Mean Squared Error (MSE) and the R-squared (R^2) score, both of which are standard metrics for assessing regression models.

> The MSE quantifies the average squared difference between the predicted values and the actual values, providing a measure of the model's prediction accuracy. It is calculated using the formula.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- n is the number of observations.
- y^i represent the predicted value of the i-th observation.
- yi denotes the actual value for the i-th observation.

Additionally, the R^2 score, which represents the proportion of variance in the dependent variable that is predictable from the independent variables, was used. The R^2 score is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:

- $\bar{y}$ is the mean of the actual values

For our dataset, the following results were obtained:

- **Total instances used**: 212,562
- **Training set size**: 170,051
- **Test set size**: 42,511

The dataset was split into training and test sets using an 80-20 ratio. This split was performed using the `train_test_split` function from the scikit-learn library, ensuring that 80% of the data was used for training the model, while 20% was reserved for testing. This random split helps to prevent bias and ensures a more robust evaluation of the model's performance.

The performance metrics for our model are as follows:

- **Training Mean Squared Error (MSE)**: 324,588,869,919.25
- **Testing Mean Squared Error (MSE)**: 446,671,197,070.33
- **Training R^2 Score**: 0.9155
- **Testing R^2 Score**: 0.9320

The relatively low MSE values and high R^2 scores on both the training and test sets indicate that our model performs well in predicting the export values of crop products, with a strong ability to generalize to unseen data. These results demonstrate that the model effectively captures the underlying patterns in the data, providing accurate forecasts for future export values.

MLP Model

Model Description for MLP Model

To forecast agricultural export values, we developed a machine learning model utilizing a Multilayer Perceptron Regressor (MLPRegressor). The model was trained on a cleaned dataset, focusing on features related to food balances, consumer prices, crop production, and food trade. These features were selected due to their direct relevance to agricultural export patterns and their potential influence on future export values.

Feature engineering was employed to enhance the model's ability to capture complex relationships within the data. Polynomial interaction terms were created to account for the potential interplay between various features. The target variable, "Export Value ($)," was shifted three years forward to align with the input features, enabling the model to learn the relationship between current agricultural conditions and future export outcomes. Rows with missing values resulting from this temporal shift were removed to maintain data integrity.

Following data preprocessing, which included feature scaling to ensure optimal model convergence, the dataset was divided into training and testing sets. The MLPRegressor

model, configured with two hidden layers (64 and 32 neurons), was then trained using the training data. The model employed the ReLU activation function for hidden layers and the Adam optimization algorithm for weight updates. The maximum number of iterations for training was set to 500.

The MLPRegressor model is defined as follows:

MLPRegressor(hidden_layer_sizes=(64,32), activation='relu',solver='adam',max_iter1000)

The model's performance metrics on the test set indicated strong performance, with a Testing Mean Squared Error (MSE) of 446,671,197,070.33 and a Testing R^2 Score of 0.9320. These results demonstrate that the model effectively captures the underlying patterns in the data, providing accurate forecasts for future export values.

Activation Function

The hidden layers employ the Rectified Linear Unit (ReLU) activation function, defined as:

$$\text{ReLU}(x)=\max(0, x)$$

ReLU is chosen due to its ability to introduce non-linearity into the model and its computational efficiency. By setting negative values to zero, ReLU helps the model learn complex patterns without causing the vanishing gradient problem that is commonly encountered with other activation functions like sigmoid or tanh.

Loss Function

The model was trained using the Mean Squared Error (MSE) loss function, which is minimized during training to boost model accuracy. MSE is perfect for regression tasks because it directly penalizes large errors more than small ones, increasing the likelihood that the model will produce accurate predictions. The MSE is calculated as

$$\text{Loss} = \frac{1}{n}\sum(y_i - \hat{y_i})^2$$

Output Layer

A linear activation function is used in the output layer for regression problems where the target variable may have any real value. Because of the linear activation, the model can predict a wide range of values without being restricted to a specific range.

Optimizer

The Adam optimizer was chosen for training due to its adaptive learning rate and efficient handling of sparse gradients. Adam combines the advantages of two additional stochastic gradient descent extensions, AdaGrad and RMSProp. It computes distinct adaptive learning rates for different parameters, which performs well in scenarios with large datasets and noisy gradients.

In summary, our MLP model shows promising results in predicting agricultural export values, with high accuracy and reliability as evidenced by the performance metrics. Further refinement and exploration of additional features could enhance the model's predictive capabilities even further.

Selected Features

For the model, a total of 11 features were chosen based on their relevance to predicting agricultural export values. These features were selected due to their significance in previous research and their strong correlation with the target variable. The selected features are:

1. **Area**: Represents the geographic region where crops are grown. This feature captures regional differences in climate, soil, and other factors that influence crop production.
2. **Year**: Indicates the year of the data, helping the model understand temporal trends and patterns over time.
3. **Yield**: Measures the agricultural productivity per unit of land. Higher yields generally lead to higher production volumes and potentially higher export values.
4. **Agricultural Use_x**: Refers to the land area used for agriculture, providing insight into the scale of agricultural activities.
5. **Emissions (CO2)**: Represents the carbon dioxide emissions from agricultural activities, which can impact crop health and production.
6. **Emissions (N2O)**: Represents the nitrous oxide emissions from agricultural activities, another factor affecting crop health and production.
7. **Export Quantity**: Measures the historical export volume of various crops, serving as a direct indicator of export performance.
8. **Import Quantity**: Indicates the volume of imported agricultural products, which helps understand the balance between domestic production and consumption.
9. **Temperature Change**: Captures changes in land temperature, which can affect crop yields and agricultural productivity.
10. **Crops total (Emissions CH4)**: Represents methane emissions from crops, which can influence the environmental impact of agricultural practices.

11. **Crops total (Emissions N2O)**: Represents nitrous oxide emissions from crops, providing additional context on environmental factors affecting agriculture.

Pre-processing

Step in Pre-processing

Pre-processing techniques played a crucial role in preparing the data for the MLP model. These steps ensured that the dataset was clean, comprehensive, and suitable for training. The following details outline the preprocessing techniques applied:

Data Cleaning

We began by eliminating unnecessary or redundant information from certain columns across all 13 CSV files. Columns with high correlation to others or those with minimal impact on the prediction task were removed to reduce the dataset's size and complexity, ensuring only relevant information remained.

Merging Data

Thirteen CSV files were merged to create a comprehensive dataset for modeling. This step involved aligning data from multiple sources, ensuring consistency by matching timestamps and key identifiers. The CSV files used included:

1. **Consumer Prices Indicators**
2. **Crops Production Indicators**
3. **Emissions**
4. **Employment**
5. **Exchange Rate**
6. **Fertilizers Use**
7. **Food Balances Indicators**
8. **Food Security Indicators**
9. **Food Trade Indicators**
10. **Foreign Direct Investment**
11. **Land Temperature Change**
12. **Land Use**
13. **Pesticides Use**

Selection of Features

Features were chosen based on their relevance to the prediction task and their correlation with the target variable. Correlation analysis using heatmaps helped identify and retain features that had a strong relationship with the target variable, reducing dimensionality and focusing on the most important features.

Managing Missing Values

To handle missing values, imputation techniques such as mean or median imputation were applied. This ensured that the dataset was complete and suitable for model training, improving the overall quality of the data and preventing issues during the training process.

Standard Scaling

Standard scaling was applied to the features to ensure they had a mean of zero and a standard deviation of one. This step was crucial to guarantee that all features contributed equally during the model training process and to prevent features with larger scales from dominating the learning process. The standard scaling formula used was:

z=x−μσz = \frac{x - \mu}{\sigma}z=σx−μ

Where:

- xxx is the feature value
- μ\muμ is the mean
- σ\sigmaσ is the standard deviation

Establishment of the Target Variable

A new column named `export_value_3yr` was created to represent the export value three years into the future. This target variable was essential for the predictive task, allowing the model to learn from historical data and make future forecasts.

Polynomial Features

Polynomial features were created to capture potential non-linear relationships between the features and the target variable. This involved generating higher-order terms and interaction terms for the selected features, helping the model understand complex interactions that linear features alone might not capture.

Hyperparameter Tuning

GridSearchCV was employed to fine-tune the MLP model's hyperparameters. A range of values for hyperparameters such as the number of neurons, learning rate, and regularization strength were defined, and the model's performance was evaluated for each combination. Hyperparameter tuning was crucial to maximize model performance and ensure the chosen parameters generalized well to new data.

Model Training

The MLP model was trained using the processed dataset. The training involved feeding the input features to the model, updating the model weights with the Adam optimizer, and computing the loss using the Mean Squared Error (MSE) loss function. The model learned from the data over multiple epochs.

Evaluation

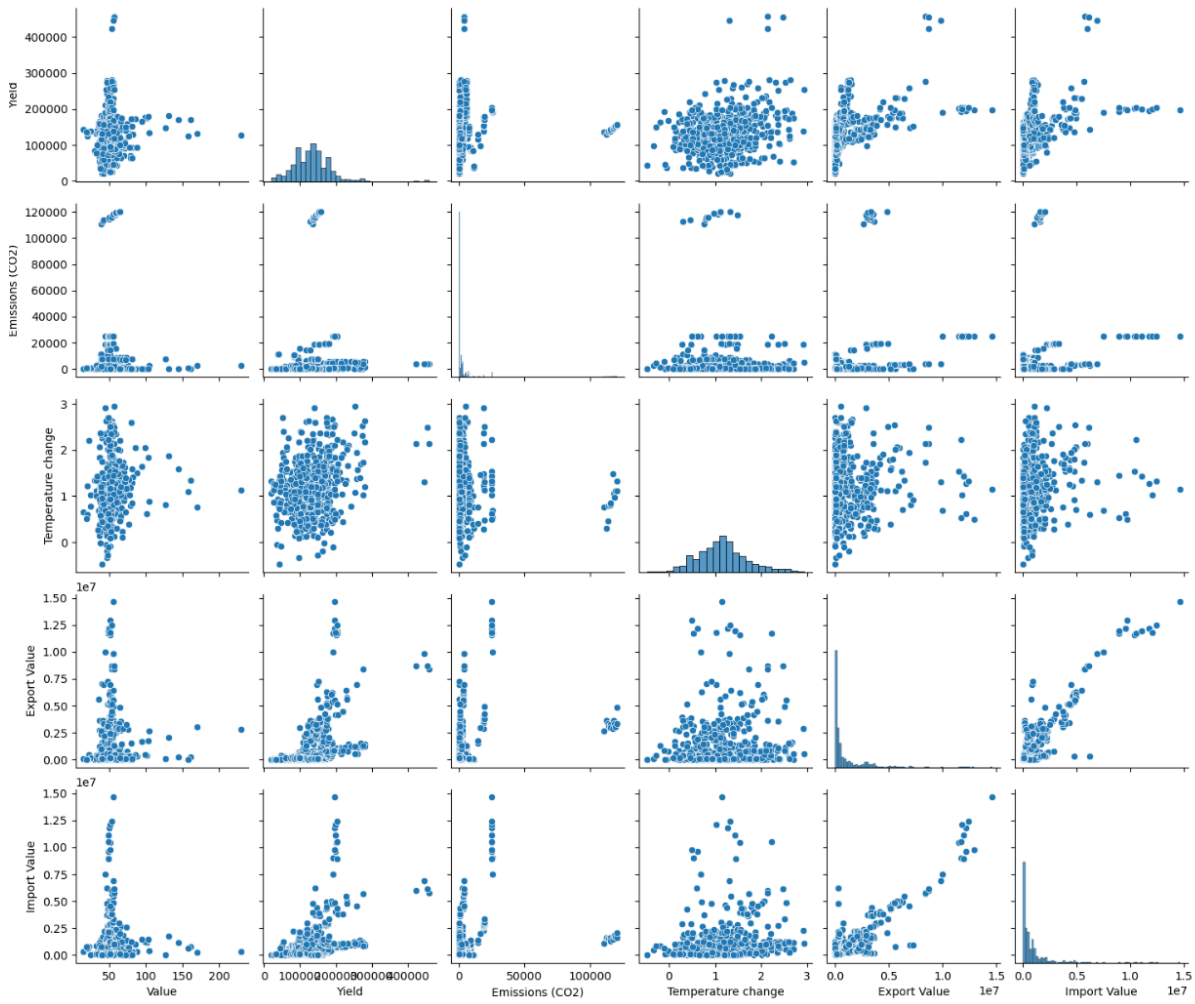
The model's performance was evaluated using the MSE and R^2 metrics. Performance metrics were calculated, results were interpreted, and the predicted values were compared to actual values in the test set. This evaluation phase was critical for assessing the model's accuracy and its applicability in predicting future export values.

This comprehensive pre-processing ensured that the dataset was well-prepared for training the MLP model, ultimately leading to a robust and accurate predictive model for agricultural export value

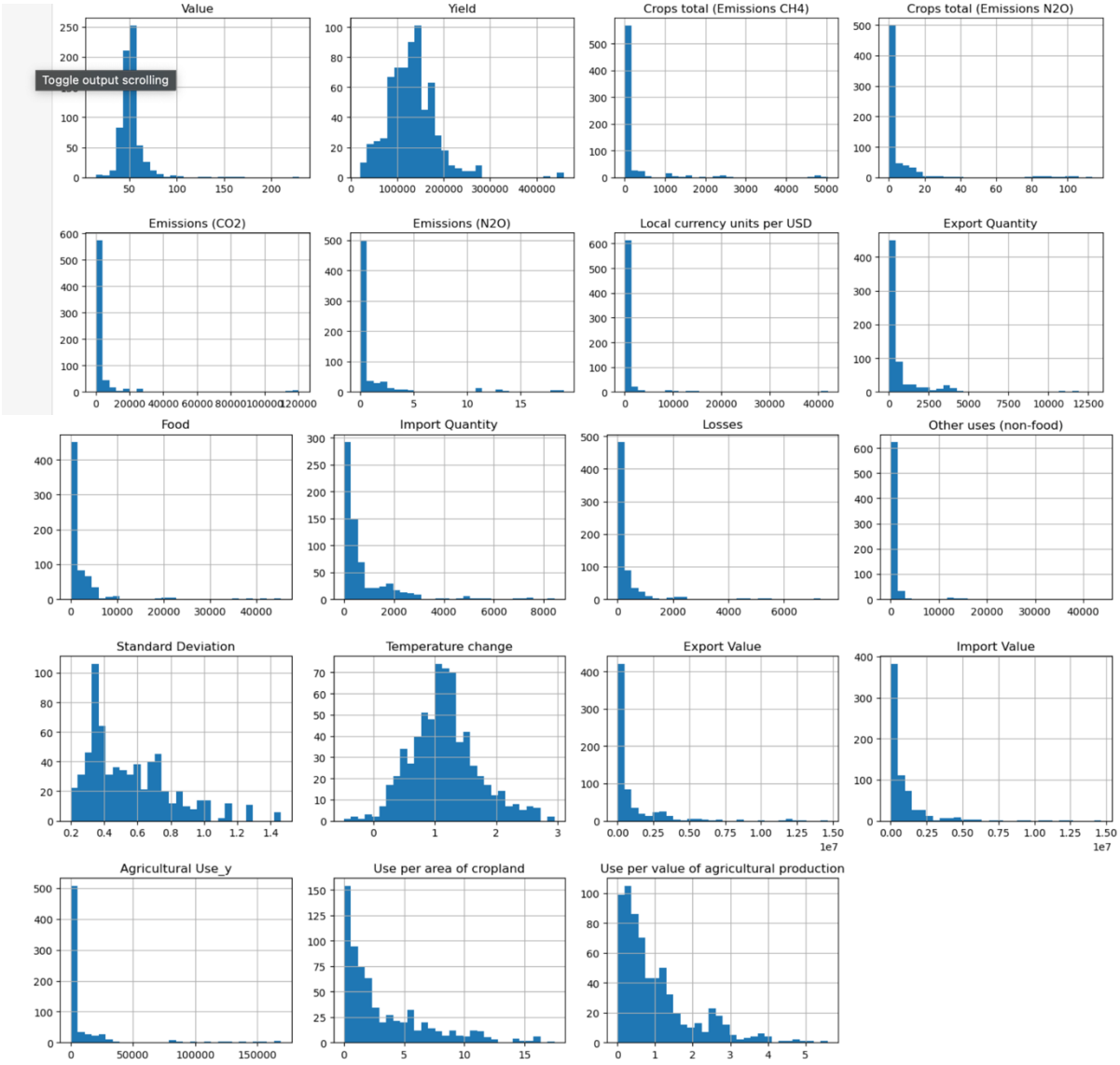
Analysis and Visualization

To view and evaluate the predictions made by the model:

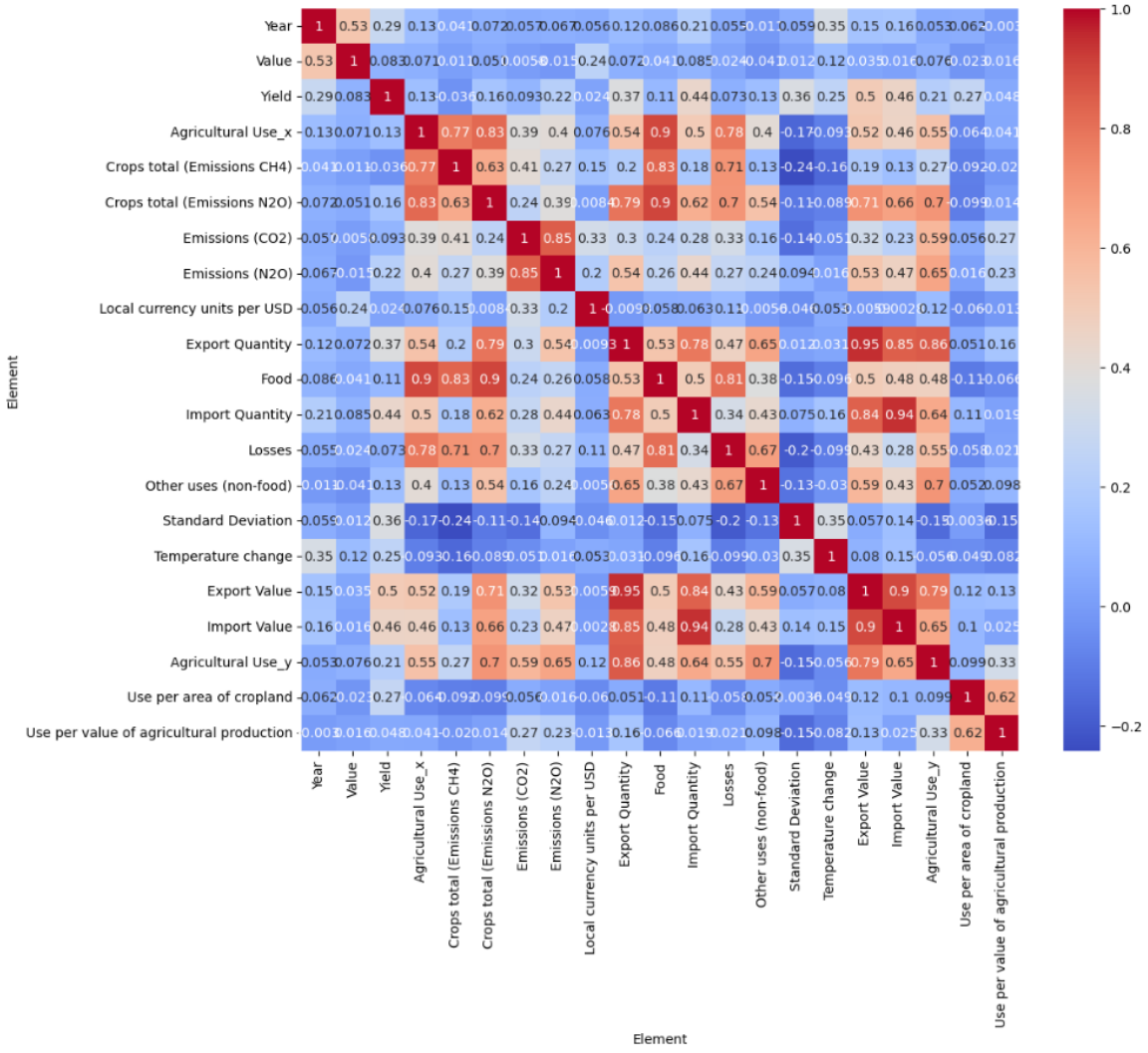
- Scatter Plots: Scatter plots were created to compare actual values with expected values in order to help identify any trends or inconsistencies.



- Histograms: The distribution of prediction errors could be assessed by plotting the residuals, or errors, as histograms.



The correlation heatmap (shown in the image) helped in identifying strong correlations between features and the target variable, further guiding the feature selection process to ensure the most relevant features were included in the model.


Element
Year
Value
Yield
Agricultural Use_x
Crops total (Emissions CH4)
Crops total (Emissions N2O)
Emissions (CO2)
Emissions (N2O)
Local currency units per USD
Export Quantity
Food
Import Quantity
Losses
Other uses (non-food)
Standard Deviation
Temperature change
Export Value
Import Value
Agricultural Use_y
Use per area of cropland
Use per value of agricultural production
Element
1.0
0.8
0.6
0.4
0.2
0.0
−0.2