Predicting the Imports using Machine Learning

Son Tran

April 15, 2025

1. Introduction

Understanding what factors affect imports is crucial for developing good trade policies. As countries rely more on international trade, a good trade policy can greatly determine the growth of a country's economy. In this project, I aim to determine the feature variables and models that best predict the monthly import value of a country.

2. Problem Formulation

For this project, the label will be ln_imports, which is the monthly natural log of imports, measured in millions of USD. There will be two sets of features. The first set will be month-to-month official exchange rate percentage change (measured as local currency per USD), CPI price, ln_labor, annual GDP growth rate, annual GDP per capita growth rate, imports share of GDP, and unemployment rate. The second set will omit CPI price, annual GDP growth rate, annual GDP per capita growth rate, unemployment rate, and instead include annual GDP. Table 1 shows the variables in detail.

The data is collected from two datasets, the Global Economic Monitor (GEM) and World Development Index (WDI) from the World Bank Group. These indexes obtain data primarily through official sources like national statistical agencies. Therefore, inconsistencies are common for many countries. Data from the GEM—which contains variables such as imports, exports, exchange rates—are monthly, while data from the WDI—which contains labor, GDP, exports and imports share of GDP statistics—are annual. The two datasets are merged by assigning the same yearly observations to all months of that year.

The final dataset contains 2226 observations. From around 200 countries listed in the two indexes, only 21 countries have reliable data that can be used for this study. The time period is between February 2010 and December 2018. Most of the countries in this data are either developed, or are developing with a large economy. The dataset is split chronologically into the training, validation, and test set with a 70:15:15 ratio. This ensures that the model is trained on earlier data and evaluated on future data.

Table 1: Label and Features Definition

Variable	Definition
Label	
ln_imports	Difference of natural log of imports in million USD (Monthly).
Features	
official_exchange_rate_percent_change	Monthly percent change in official exchange rate (LCU/USD).
CPI_percent_change	Monthly inflation rate.
ln_labor	Natural log of labor force (Annual).
GDP_annual_growth	Annual GDP growth rate.
GDP_per_capita_annual_growth	Annual GDP per capita growth rate.
export/import_share_of_GDP	Percent export/import share of GDP (Annual).
unemployment_rate	Unemployment rate (% of labor force) (Annual).
GDP_annual	Annual GDP.

3. Approaches

For the approach, I will be using three different machine learning algorithms: linear regression, ridge regression, and decision tree regression. The hyperparameter of ridge regression will be alpha, representing the regularization strength, while for decision tree regression, it will be the depth. The hyperparameters will be tuned to find the best values, and then results of the two sets of features are compared.

The evaluation of success will be Mean-Squared Loss (MSE) of the test set. This metric is not the real goal of the task — which is to accurately predict future imports — but serve as reasonable approximations. MSE punishes big mistakes more and gives an idea of how far off the predictions are on average, making it useful for checking how accurate the models' predictions are.

4. Results

The result table shows that for linear and ridge regression, the second feature set is slightly better at predicting imports than the first feature set. For decision tree regression, the model with the first feature set performs better.

	First Feature Set	Second Feature Set
Linear Regression	Test MSE: 0.35204161438841824	Test MSE: 0.32158500390490197
Ridge Regression	Best alpha found: 100.0 Test MSE: 0.35151371318541524	Best alpha found: 0.01 Test MSE: 0.321600479690388
Decision Tree Regression	Best max_depth found: 7 Test MSE: 0.08421464224688598	Best max_depth found: 9 Test MSE: 0.09616458988658386

Based on the results from figure 1, when using the first feature set, the validation MSE from the ridge regression starts at around 0.417, and gradually decreases down to around 0.411 at $\lambda = 100$. However, when looking at the figure 2, using the second feature set allows the validation MSE to start at a much lower value, at less than 0.34. Increasing λ only leads to higher MSE. This suggests that the first feature set is noisy and may contain redundant features. This may cause the model to overfit, and increasing the λ helps reduce this. In addition, the second feature set, despite containing less features, allows for more accurate predictions overall.

Interestingly, the validation MSE for the model using the first feature set is notably higher than the test MSE. This suggests that the validation set may represent a time period with different characteristics than the train set and test set. However using the second feature set removes this gap, suggesting that it can capture these characteristics better.

Figure 1: Ridge Regression: Validation MSE vs λ (First Feature Set)

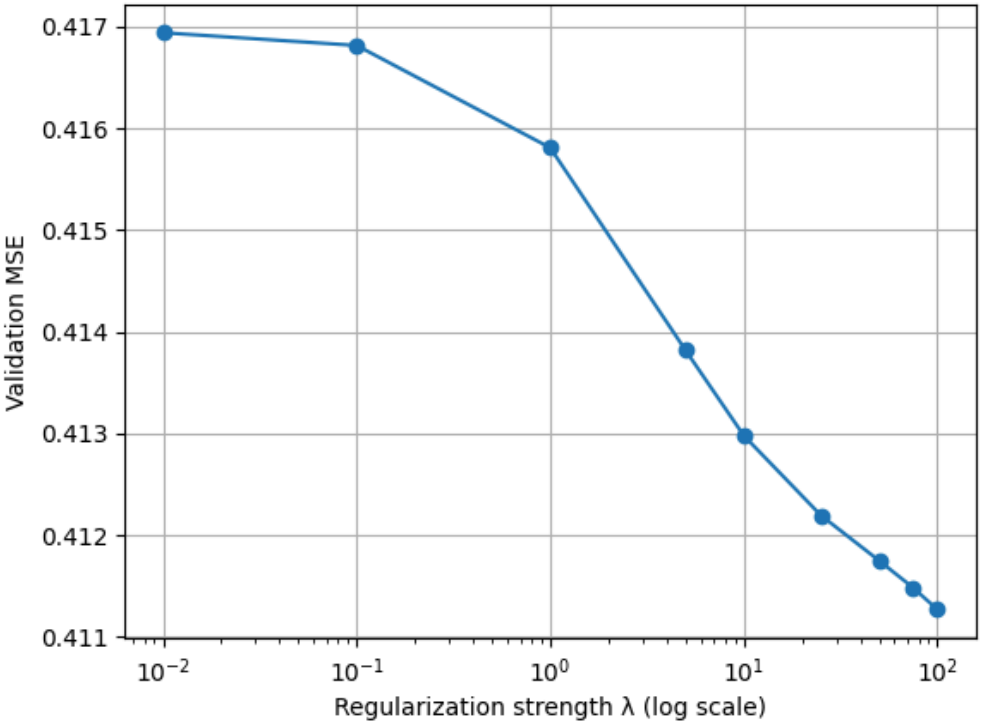
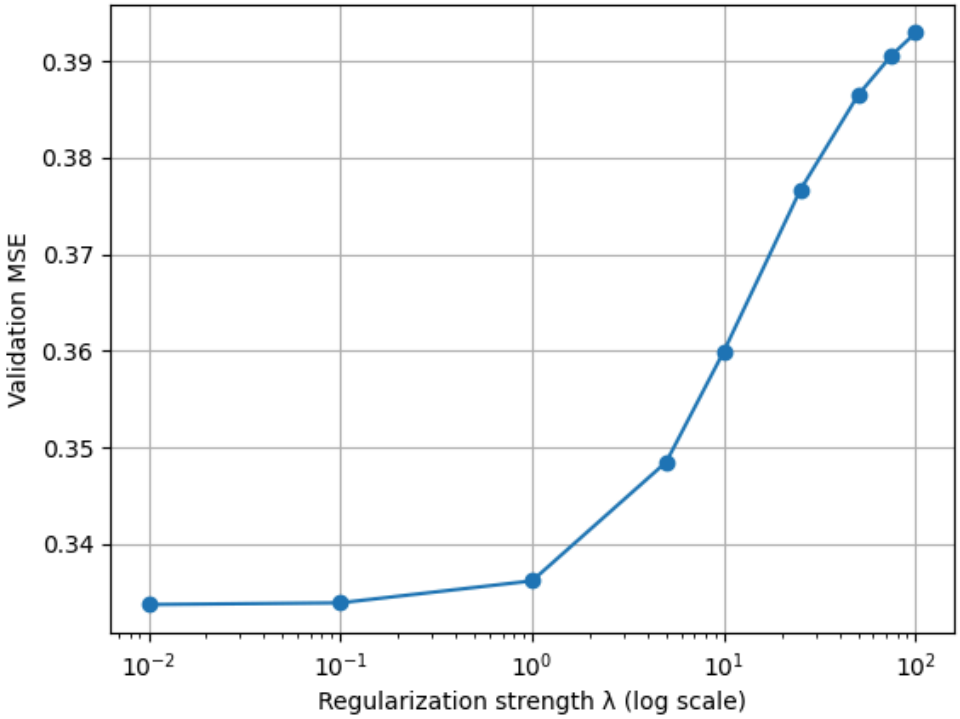


Figure 2: Ridge Regression: Validation MSE vs λ (Second Feature Set)



Figures 3 and 4 show that decision tree regression predicts the value of imports better than linear or ridge regression overall. However, the models of the first feature set and the second feature set show somewhat similar validation results.

Figure 3: Decision Tree: Validation MSE vs Max Depth (First Feature Set)

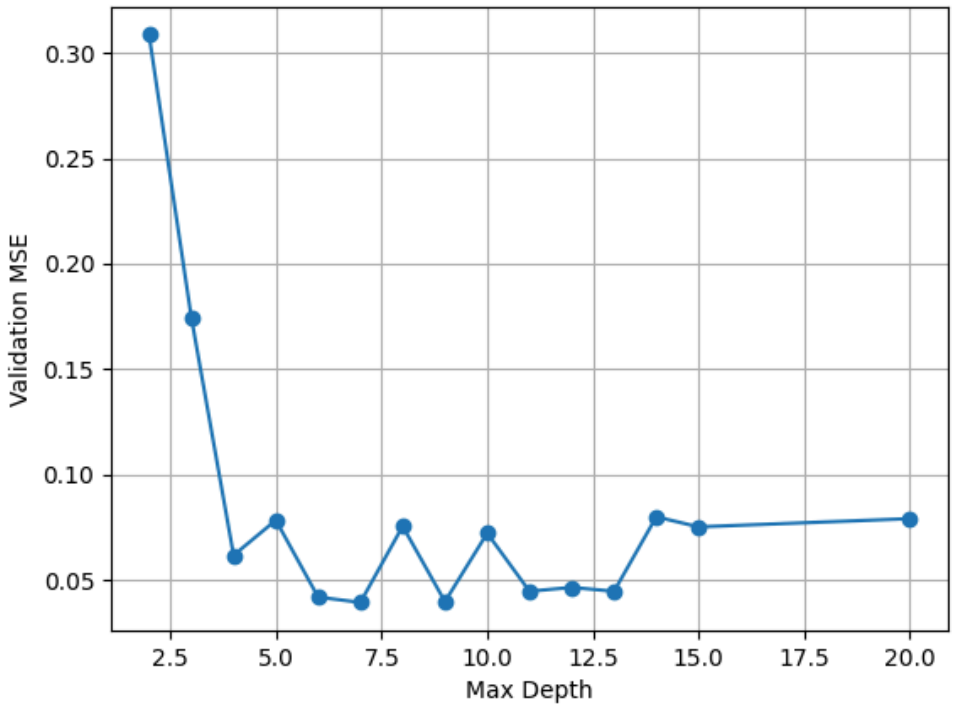
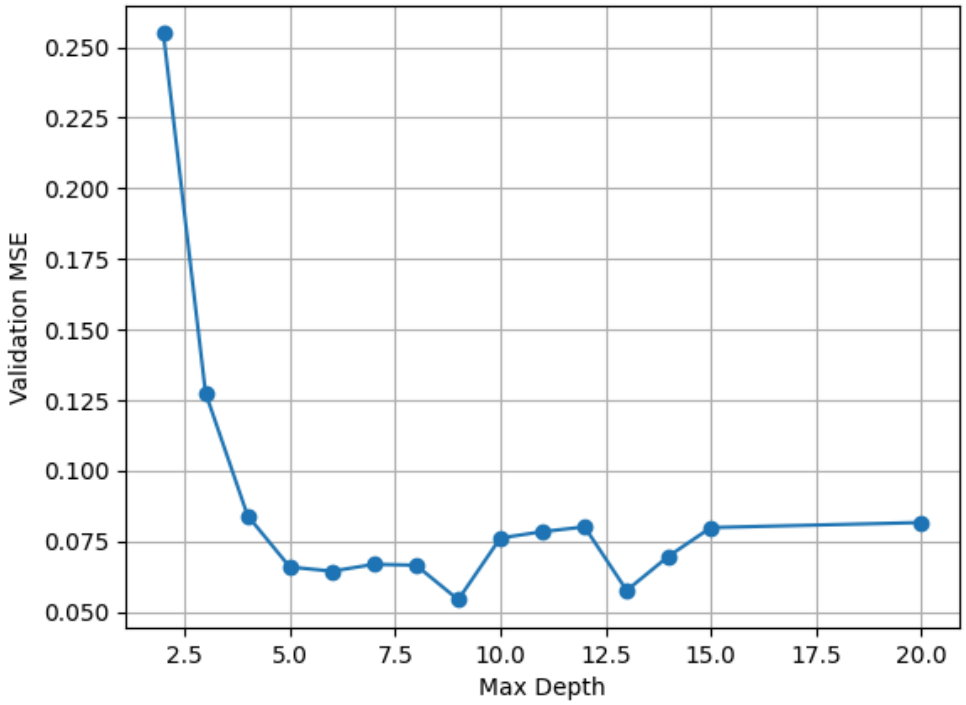


Figure 4: Decision Tree: Validation MSE vs Max Depth (Second Feature Set)



Overall, the linear and ridge regression shows better performance with the second feature set than the first feature set. The second feature set appears to include features that better correlate with the label ln_imports. Decision tree regression outperforms both linear regression and ridge regression with both datasets. This suggests that the relationship between imports and the features are non-linear.

Sources

World Bank. (2025). *Global Economic Monitor (GEM)*. World Bank DataBank.

World Bank. (2025). *World Development Index (WDI)*. World Bank DataBank.