Juanita Santofimio
Js5972

Predicting Steel Prices through Analyzing the Impact of Conuntry-Specific Supply Changes after 2018 Tariffs

GITHUB LINK: https://github.com/js5972/aml_project1.git

Background:

With the rising uncertainty regarding upcoming tariffs from the current Trump administration, it has become imperative for U.S. domestic manufacturers to strengthen supply chains. While there is widespread agreement that tariffs will increase import prices, this is accompanied by periods of high volatility that affect manufacturing. A particular example is the case of steel, which is an input to most of the manufacturing sector. Tariffs on steel, aluminum, and derivative goods currently account for $2.7 billion of the $79 billion in tariffs, based on initial import values. This comes from the tariffs on steel imposed by the previous Trump administration in 2018 under Section 232, citing national security concerns. Recently, in February 2025, Trump announced the ending of steel exemptions totaling $29 billion, expanding derivatives by $44 billion under chapters 73 & 76, plus metals content of an additional $100 billion, on top of the existing 25% rate imposed in 2018. This took effect on March 12, 2025.

Since, a key factor to determining prices and costs of production is suppl, this project conducts exploratory analysis to see if the previous month value of supply of steel can be used to predict next month prices, in particular noting the difference between pre and post 2018 tariffs and accounting for subsequent country specific agreements

Target User: Procurement Manager:

Roles:

- Source and purchase materials, goods, and services for the company.
- Negotiate contracts, maintain supplier relationships, and ensure timely delivery.
- Manage budgets, inventory, and procurement processes.

Impact of Steel price volatility:

- Unpredictable production costs.
- Disruptions in supply chains and cause delays in steel shipments.

How Your Work Benefits from Understanding Steel Prices:

- Budgeting**:** factor in expected of price increases.
- Supplier Negotiations**:** secure better terms based on updated price expectations.
- Risk Management: make sourcing decisions based on available information on dependency on specific countries as sources of steel.
- Planning: scheduling orders based on expected price.

Objective:

1. The project aims to analyze the impact of tariffs, specifically those imposed by the Trump administration around 2018, on prices od steel in the U.S.
2. The primary focus is to study how steel trade volumes change before and after tariffs for different countries and how steel prices react accordingly to such supply changes.

Features:

A lot of consideration was put into thinking what country specific features for both the US and trading partners were reasonable to identify salient countries in steel price prediction, while preserving model interpretability. Some of these features were considered with a different project objective in mind and were therefore discarded after (refer to last section).

Discarded features:

- Manufacturing PMI: considered as a measure of domestic demand to account for cyclical and industry specific factors. However, after a test run, it became evident that the collinearity between this variable and steel prices, would overwhelm any other variable in the model.
- Trade partners macro economic data: considered as a way to account for the trade environment and international economic conditions but discarded because any country specific information would be priced into import value, and instead it would reduce model interpretability.
- Distance: considered as a feature to measure similarity when the project objective was about clustering, but irrelevant when trying to predict based of import volumes.
- Import duties: considered as a way to explicitly factor in tariffs, but it is really just a function of import size and wether it was pre or post tariffs, which is already measured by the corresponding variables.

Final selection:

- Imports volume per country per month between 2015 and 2021 (3 years before and after tariffs)
- Trade policy including when the steel tariffs were announced and effective, and the different dates at which countries negotiated agreements like exclutions and quotas
- Steel prices using the NYSE Steel Index

Data Preparation Choices:

- **Missing values:** In terms of import size, for many countries there were multiple missing values since there are months in which there is no trade. Depending on the proportion of missing values, some countries had those months filled in with zeros and other countries for which there was trade only in very few occasions were simply dropped.
    - The specific dropping criteria were based on the count of months for which % change in import value was missing. This was the case because the project objective was different at the time of cleaning the data (refer to last section). Data was dropped when there were big desserts of inactivity, as change in trade volume would be very high (from 0 to millions value or vice versa), which creates noise and is not informative about supply. Countries for which there was activity in less than 25% of reported months (which was already low for most) were dropped.
      This decision remains equally valid for the final model in terms of noise reduction.
- **Scaling:** due to the wide range of import values, from zero to high millions, scaling was necessary. Tariff related variables were binary and therefore needed no scaling.

Model Choices:

- Supervised learning since we are training on observed, correct values.
- Regression models given steel prices are continuous.
- **OLS-PCA:** reduce dimensionality to simplify the model because having one beta per country means most will be insignificant.
- **Lasso:** since most countries are probably going to be very poor predictors or insignificant indicators of steel prices, we can benefit from bringing the coefficients down to zero through L1 regularization.
- **Ridge:** COVID is an external shock that also affected all countries after the tariffs were imposed. Therefore, using L2 regularization will reduce the impact of this collinearity while maintaining any signals each country contributes.
- **Decision Tree:** provides insight on the importance of each country and the extent to which the size of each import matters in deetermining the prices of steel.
- **Random Forest:** reduces the risk of overfitting from a single tree.
- **Gradient Boost:** increases the prediction accuracy of prices by shrinking the contribution of each country, modeling based on little combinations of all rather than prioritizing the most salient ones.

Evaluation metric:

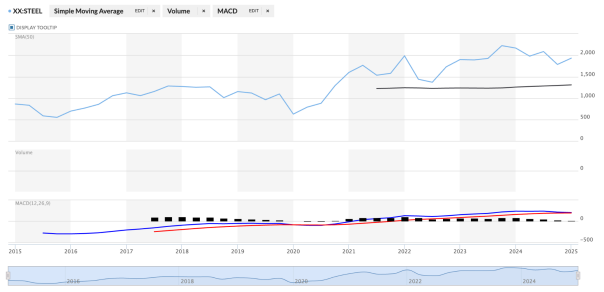
- **MSE:** penalizes larger errors more capturing the effect that unexpectedly high/low levels of supply can impact steel prices more. Otherwise observations tend to be clustered around the mean and vary more for cyclical reasons rather than supply related factors.

Results and Analysis::

The results across models suggest a lack of consistent predictors, as very little overlap exists between the features selected by different models. This inconsistency implies that the relationship between steel imports and prices may be more random or influenced by external factors rather than driven by strong, identifiable predictors.

Ridge	Random Forest	Decision Tree	Lasso	PCA	Gradient Boost
HONDURAS	OMAN	HONDURAS	HONDURAS	MEXICO	HONDURAS
DENMARK	ECUADOR	MEXICO	DENMARK	OMAN	ECUADOR
COSTA RICA	BULGARIA	BAHRAIN	BELARUS	ECUADOR	BULGARIA
BELARUS	COLOMBIA	GERMANY	COSTA RICA	FRANCE	DOMINICAN REPUBLIC
GEORGIA	DOMICAN REPUBLIC	GUATEMALA	DOMICAN REPUBLIC	GERMANY	

One notable observation is the frequent appearance of Honduras as a predictor across multiple models. This is likely due to collinearity with other variables or an external factor, rather than Honduras being a true driver of steel price changes. The most plausible explanation is similarity in economic cycles.

NYSE American Steel Index Advanced Charts | MarketWatch	Honduras Steel Exports
	

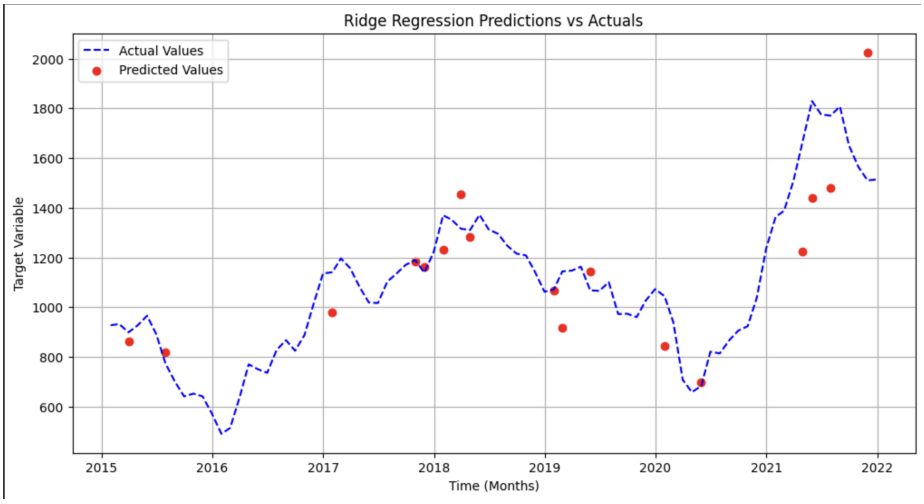
Model Ranking and Performance

The models ranked as follows:

1. **Ridge Regression –** Best performing. Most likely because, as mentioned before, can capture signals from across all factors, but applying shrinkage to remain generalizable. In the context of imports this makes sense as all contries contribute to total supply, but some are too small to provide valuable information.
2. **Random Forest –** Performed better than the tree because it can gather more information from different potential splits.
3. **Decision Tree –** When running with multiple different random states, the optimal tree parameters varied significantly, hinting that the results are not necessarily very generalizable. There the chosen tree is most likely overfitted.
4. **Lasso –** bringing about 65% of the coefficients to zero could have resulted in some information loss.
5. **OLS-PCA –** The top countries in the main components on the PCA were very different from the rest of the models. This means the phenomena that explain movements in import volumes are not useful to predict price changes.
6. **Gradient Boost –** Worst performance, most likely due to an error in the pipeline because I would expect better performance at the very minimum than the OLS-PCA. Hard to debug due to extended running time.

Ridge Regression Analysis

The Ridge regression model captured the general trend of steel prices but lacked precision in individual predictions. The graph suggests that while Ridge avoids extreme errors, it does not provide reliable predictions, reinforcing the idea that external or unaccounted factors play a dominant role in steel price movements.



Overall, the models demonstrate that predicting steel prices based solely on import values is highly uncertain due to collinearity, external factors, and market complexity. While Ridge Regression provided the most stable results, the predictive power of these models remains limited, suggesting that additional economic, geopolitical, and demand-side factors may be necessary for more accurate forecasting.

Learning points:

Better Problem Definition from the Start: It's crucial to not only define a very precise question, but also to clearly outline how the data should look from the beginning. Focusing on this helps avoid the trap of collecting numerous features without a clear strategy, as I initially did.

Clustering and Port Activity Analysis: My original idea was to use clustering to identify collinearity in port activity and anticipate fluctuations in activity in order for a port manager to better prepare personnel, machinery, and capacity. However, I spent considerable time downloading, calculating, and cleaning features, only to realize that the larger picture was much more complex and unclear than I initially thought.

Looking forward to applying these lessons to the final project, with a more focused and strategic approach.