Foreign Exchange Rates: Predicting the Euro with Leading Variables

Riley D. Fiske

Concordia College

DATA-318: Data Mining

Dr. Gregory Tanner

April 28, 2022

Foreign Exchange Rates: Predicting the Euro with Leading Variables

Investopedia defines globalization in economic terms to be "the interdependence of nations around the globe fostered through free trade."[1] The world we live in today is a world of globalization, where nations trade among each other to gain access to materials, goods, or services they would have otherwise had to develop on their own which is more expensive. One problem that needed to be solved when this began was that since different countries use different currencies, how can countries buy foreign goods. One way is to buy foreign currency with an agreed upon conversion rate. This conversion rate depends on the demand of goods or services in either country, economic and foreign policies of the governments, the political climate of the time, and the price of gold.

A currency is considered stronger when it takes more of another currency to get less of that currency. For example, on April 22nd, 2022, it took 1 US Dollar to buy 0.9244707 Euros, meaning US$1=€0.9244707 and the Euro is stronger than the US Dollar.

This statistical project using data mining techniques sought to predict the exchange rate of the Euro relative to the US Dollar using other foreign currency exchange rates relative to the US Dollar from two previous days. There are 33 currencies listed alongside the date for which the rates correspond to, however the Turkish Lira (TRL) was filtered out as in 2005 it was transformed into the new Turkish Lira (TRY). These currencies were then lagged by two days, and the daily exchange rates filtered out so the rate of the Euro is being predicted using previous data as that is how it would be predicted in the real world.

[1] Jason Fernando, *Globalization*, Investopedia.com

Initially the models created using all 32 currencies for two days (64 predictors) was quite accurate, but that was because certain currencies have locked themselves to the Euro's exchange rate due to how powerful of a currency it is, and due to the countries doing frequent business with the EU. All other European currencies were filtered out as a result, as well as the Canadian, Australian, and New Zealand Dollars since they are part of the British commonwealth, to eliminate as much leakage as possible into the model. This left 32 predictors in the dataset.

The created model can be used to make predictions of what the exchange rate of the Euro will be on any given day given the two previous days exchange rate for other currencies, which is quite powerful in the realm of investing and observing patterns to know when to buy or sell shares. The model is simple enough as well to be used by those with little understanding of data analytics as the predictors are readily available online.

Finding the Data

The data used for analysis was collected by the Humanitarian Data Exchange (HDX)[2], which is managed by the United Nations Office for the Coordination of Humanitarian Affairs. The dataset pulls from the European Central Bank and converts rates to USD, which is updated daily and dates back to January 4th, 1999.

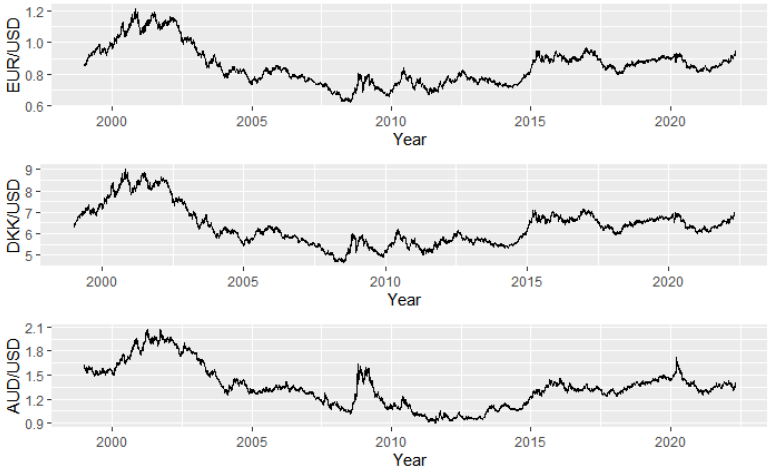
Due to the dataset being compiled by the United Nations, it is justified to say there should be few problems with it. Some potential problems with the dataset were observed, such as certain currencies not existing during certain periods or not being traded by the European Union.

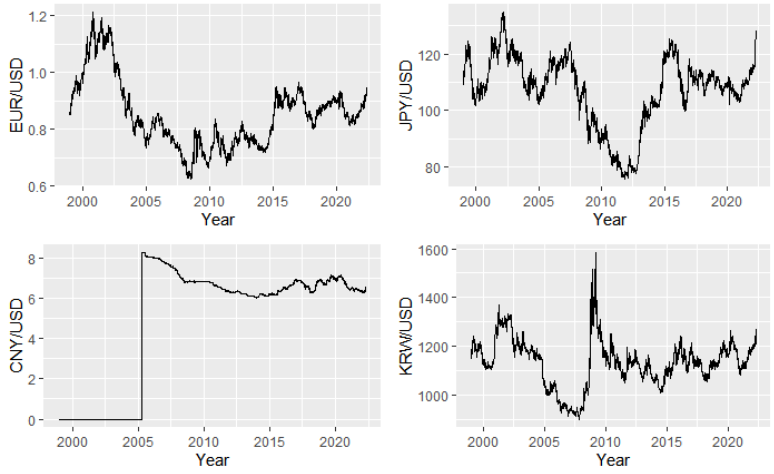
[2] Humanitarian Data Exchange. *Foreign Exchange Rates.* data.humdata.org

However, this did not appear to cause many problems as the Chinese Yuan, one such currency that falls into that category, was selected as an optimal predictor.

Exploratory Analysis

As previously discussed, certain predictors were very closely tied to the Euro and leakage was more than likely occurring. When comparing the line graphs of the Euro/USD exchange rate over time, Danish Krone/USD exchange rate over time, and the Australian Dollar/USD exchange rate over time, there are striking similarities between them. Using the coorelation function, it was determined that the Danish Krone had a 99.99% coorelation to the Euro and the Australian Dollar had a 88.40% coorelation. All non-Euro European currencies, as well as the Canadian, Australian, and New Zealand Dollars, were filtered out and currencies such as the Japanese Jen, Chinese Yuan Renminbi, and the South Korean Won remained and were used as predictors. When observing their similarity to the Euro, their trends seem to be quite destinct and more fair predictors, however the Yuan and the Won were not measured until about April 2005 and does not appear in the dataset until then. The Japanese Yen had a 56.89% coorelation to the Euro, the Chinese Yuan had a 61.83% coorelation, and the South Korean Won had a 50% coorelation.



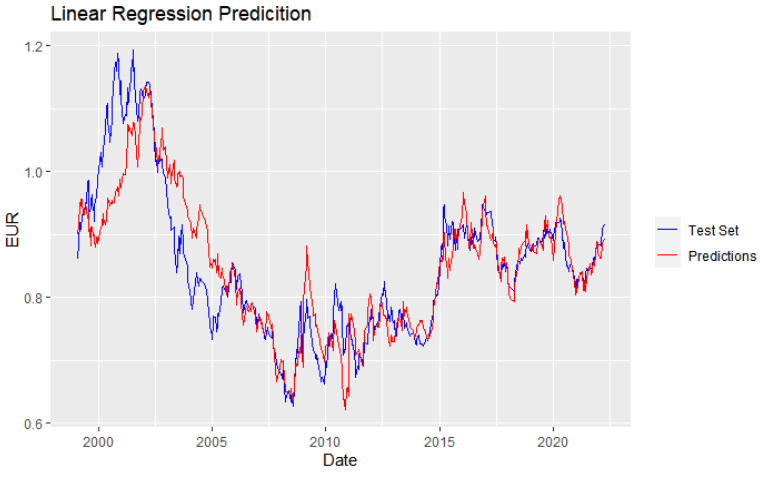


Data Modeling

Before making any predictions of exchange rates, subset selection was used to find the optimal variables to use, as well as the optimal number of variables. The Date variable was ignored in the subset selection process as date is used to graph the results of our models compared to the actual results from the test set. To select the optimal number of variables, BIC was minimized. As has been stated already, the dataset is updated daily, so as of the writing of this, the minimization selected 14 variables, which were then plugged into 4 different data models and the R^2 statistic was tracked to estimate accuracy. The data was then partitioned into a training set of 90% of observations and a test set of 10% of observations.

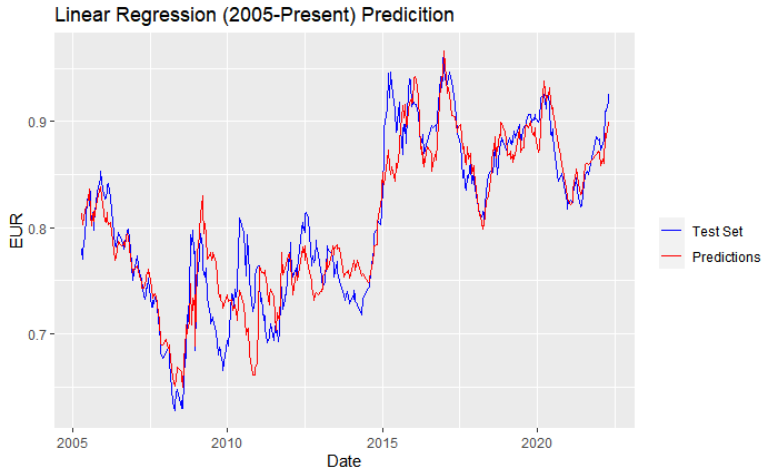
The variables selected by subset selection were Japanese Yen, Chinese Yuan, Israeli Shekel, South Korean Won, Malaysian Ringgit, Singapore Dollar, and South African Rand offset by one day and Hong Kong Dollar, Indonesian Rupiah, Mexican Peso, Malaysian Ringgit, Philippine Peso, and Thai Baht offset by two days. These variables were used for all the data models.

Linear Regression Model



Linear regression was the least accurate of the created data models but is also the most basic and interpretable. The R^2 statistic for the predictions made by this model was .7668, which can be interpreted as the model can explain about 76.7% of the variance in the data. RMSE for this model was 0.0574, which can be interpreted as the square root of the mean of the errors was

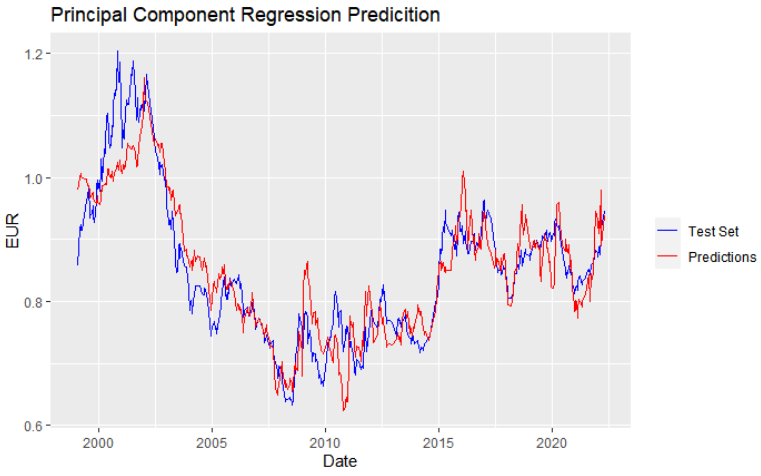
0.0574. The MAE statistic was about 0.04, which can be interpreted as the average error in either direction when estimating was 0.04.



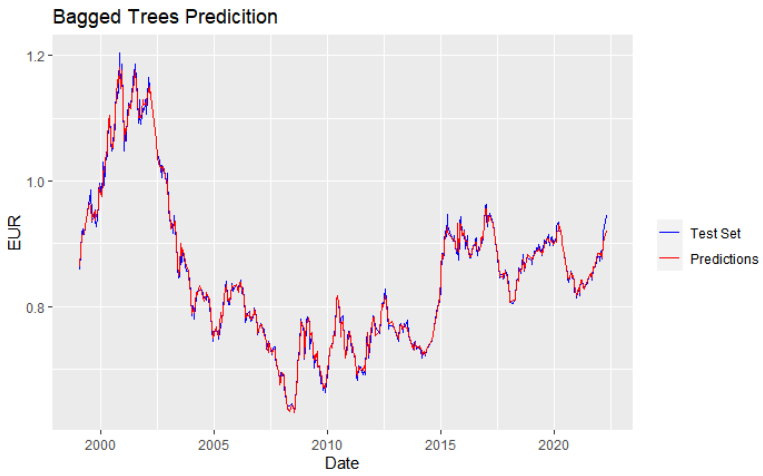
It can be seen that then trend seems to get back on track at around 2007 for this model, and one of the variables used was the Yuan which wasn't tracked until 2005, so a second linear regression model was created only observing the results after the Yuan was tracked. The R^2 statistic was 0.8756, RMSE was 0.0272, and MAE was 0.021. These three statistics indicate that this model performed better than the model for all years.

Principal Component Regression Model

Principal component regression yielded a result somewhere between the two linear models. This model selected the optimal components using cross validation, so it makes sense that it would perform close to linear regression that used our subset selection of variables. The R^2 statistic was 0.8242, RMSE was 0.0499, and MAE was 0.039.

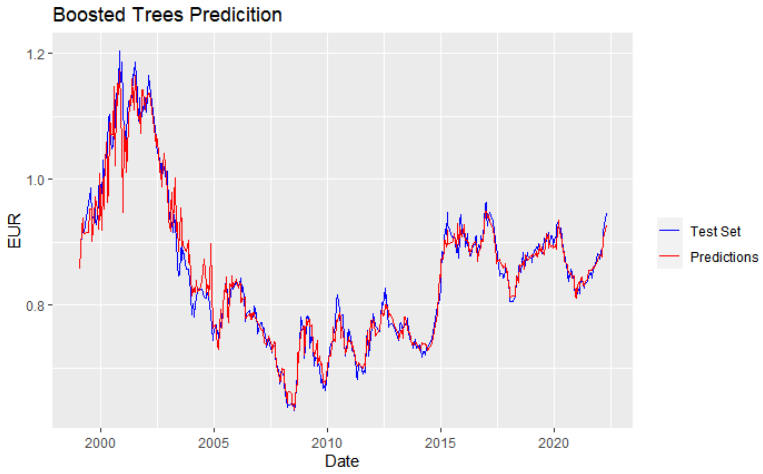


Bagged Trees Model



Bagged trees was the most accurate data model for this dataset, performing with an R^2 statistic of 0.9962, RMSE of 0.0074, and MAE of 0.0054. While this was the most accurate model, bagged trees are hard to visualize and interpret, so it acts more like a black box. When looking at the line graph, the two lines are quite similar to each other which is a clear indication of a very good fit.

Boosted Trees Model



Boosted trees performed slightly poorer than bagged trees, but much stronger than linear regression and principal component regression. Boosted trees performed at 0.963 R^2, RMSE of 0.0231, and MAE of 0.015, which is still a very good rating. On the line graph, the predicted trend is once again quite similar to the test set which indicates good fit.

Discussion

Typical error in the context of estimating foreign exchange rates should be very small and considering the fact that the most accurate predictors were removed from the model to toughen the process, it is impressive the models reached the accuracy they did. One way the accuracy could be increased is by expanding out the number of days backlog we could have as predictors

as the models can then observe trends between the days to increase accuracy most likely while still staying out of currencies with suspected leakage.

To put these models to the test, sample dates were chosen to see how close they could predict to the actual conversion rate. April 11th, 2022 had an exchange rate of 0.917 and the bagged trees model predicted 0.9195, which has an RMSE of 0.0021, while boosted trees predicted 0.9182, which has an RMSE of 0.00072. April 21st, 2022 had an exchange rate of 0.919 and the bagged trees model predicted 0.922, which has an RMSE of 0.0039, while boosted trees precited 0.9158, which has an RMSE of 0.0027. April 27th, 2022 had an exchange rate of 0.945 and the bagged trees model predicted 0.921, which has an RMSE of 0.0242, while boosted trees predicted 0.9264, which has an RMSE of 0.0184, or in comparison, a much wider gap in the mean squared error than the other days.

It appears something irregular must have happened on April 27th since the prediction was quite off, but the close predictions for the other test days shows that these models are working quite well. The writing of this is on April 27th and it appears that the current trend is the weakening of the Euro in comparison to the US Dollar. Current events in the world are more than likely causing the exchange rate of the Euro to be rising in comparison to the US Dollar, which the models are not accounting for.

Overall, these models took in a lot of observations and made some quite accurate predicting models from them that are very real world applicable. Despite there being predictors that are heavily correlated, even when filtering them out, the models perform at a considerably high accuracy. Predicting financial fluctuations are a highly studied topic, and this is merely the tip of the iceberg.

References

Alboukadel Kassambara and Fabian Mundt (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Andrea Peters and Torsten Hothorn (2021). *ipred: Improved Predictors*. R package version 0.9-12. https://CRAN.R-project.org/package=ipred

Baptiste Auguie (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

https://CRAN.R-project.org/package=gridExtra

Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2020). *gbm: Generalized Boosted Regression Models*. R package version 2.1.8. https://CRAN.R-project.org/package=gbm

Fernando, Jason. (2022) *Globalization.* Investopedia.com, https://www.investopedia.com/terms/g/globalization.asp

Fiske, Riley. (2022) DATA_318_Final_Project.RMD [Source Code] https://github.com/rdfiske17/DATA-318/blob/main/final_project/DATA_318_Final_Project.Rmd

Garrett Grolemund, Hadley Wickham (2011). *Dates and Times Made Easy with lubridate.* Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

H. Wickham (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Humanitarian Data Exchange. *Foreign Exchange Rates.* data.humdata.org, https://data.humdata.org/dataset/ecb-fx-rates?force_layout=desktop

Kristian Hovde Liland, Bjørn-Helge Mevik and Ron Wehrens (2021). *pls: Partial Least Squares and Principal Component Regression.* R package version 2.8-0. https://CRAN.R-project.org/package=pls

Masaaki Horikoshi and Yuan Tang (2016). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. https://CRAN.R-project.org/package=ggfortify

Max Kuhn (2021). *caret: Classification and Regression Training*. R package version 6.0-90. https://CRAN.R-project.org/package=caret

Terry Therneau and Beth Atkinson (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

Thomas Lumley based on Fortran code by Alan Miller (2020). *leaps: Regression Subset Selection*. R package version 3.1. https://CRAN.R-project.org/package=leaps

Wickham et al., (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686