MATH 191 TOPICS IN DATA SCIENCE: ALGORITHMS AND MATHEMATICAL FOUNDATIONS
PREDICTING EXCHANGE RATES AND GLOBAL COMMODITY PRICES

ERIC ABERBOOK, RAJYAVARDHAN PASARI

December 12, 2015

Abstract. The objective of our project is to predict prices for some important currencies and commodities. This can have several uses, but we came up with the idea because it seemed like a useful tool for traders, importers and exporters, and companies among several other groups to get an idea of what prices will look like in the near future, so that they can hedge their bets. We put together a dataset of major and exotic currencies from all around the world, as well as important commodities (a detailed list is available below). The data represents the Real Effective Exchange Rates of the currencies, and the prices of the commodities from the Chicago Mercantile Exchange (CME). To this dataset, we apply four different methods combining Linear Regression, K- Nearest Neighbors Regression, Forward Stepwise Regression and Principal Component Analysis to build four separate prediction models. After calculating our estimates, we run an error analysis on our results, and conclude that the Forward Stepwise Regression method, coupled with Linear Regression is the best prediction method to fulfill our objective.

Key words. Multiple Linear Regression, Principal Component Analysis, Stepwise Regression, k-Nearest Neighbor Regression

1. Introducing Our Dataset. We put together the Real Effective Exchange rates (REER) for twenty eight major currencies and ten important commodities, all of which are listed below. The motivation to pick the currencies is that they cover major countries that currently play, or are expected to play a significant role in the world economy. Our data represents the daily REER of these currencies over the last five years, which is tracked by Barclays. The currencies are:

NORTH AMERICA:
- US Dollar (USD)
- Canadian Dollar (CAD)
- Mexican Peso (MXN)

SOUTH AMERICA:
- Brazilian Real (BRL)
- Argentine Peso (ARS)
- Chilean Peso (CLP)
- Colombian Peso (COP)

EUROPE:
- Russian Rubel (RUB)
- The Euro (EUR)
- British Pound (GBP)
- Hungarian Forint (HUF)

ASIA:
- Chinese Yuan (CNY)
- Indian Rupee (INR)
- Philippine Peso (PHP)
- Turkish Lira (TRY)
- Taiwanese Dollar (TWD)
- Thai Bhat (THB)
- Singapore Dollar (SGD)
- Indonesian Rupee (IDR)
- Korean Won (KRW)
- Japanese Yen

OCEANIA:
- Australian Dollar (AUD)
- New Zealand Dollar (NZD)

AFRICA:
- South African Rand (ZAR)

We also picked some important commodities, which are listed below. The motivation behind picking them was that commodities play an important role in international trade, and can often influence the value of a particular currency (we see this in the case of the Russian Ruble and Oil). The prices are the daily close values of the Chicago Board of Trade (run by the Chicago Mercantile Exchange).

COMMODITIES:

- Oil
- Gold
- Aluminium
- Coffee
- Wheat
- Rice
- Livestock
- Sugar
- Cotton
- Lumber

The price points above for all of the currencies and commodities are obtained from Bloomberg.

2. Our Work.

We build four different models using variations of Linear Regression and Principal Component Analysis, and each of them is used to predict prices of instruments.

2.1. Pearson Correlation and Linear Regression.

In this method, we use the pearson correlation to determine the prediction variables (our currencies and commodities) for our response variable (again, currency or commodity). The correlation method acts as a filter; instead of using thirty seven predictor variables, we use only ten of them. The pearson correlation helps us find out the features that are the most similar to our response variable, and this filtering mechanism allows us to consider more important features in our regression model.

We use the sliding window approach to look at the data for the previous 100 days. Using this window of data, we determine the most important predictors using the pearson correlation method. We then regress the data onto the closing prices obtained for the next day of our response variable, with which we build a linear model. At this point, we are considering 101 days of data (days 1 to 100 for the predictors, days 2 to 101 for the response). We then use the coefficients to determine the value of the response for the 102nd day.

2.2. Forward Stepwise Regression and Linear Regression.

This process involves the usage of linear regression and the sliding window, but the filtering mechanism is different. Instead of using correlations, we use the forward stepwise regression method to pick our variables. This method picks the variables that explain the model best, and for 37 predictors tells us the best one we should add if we wish to increase the number of predictors we use. We use the ten best predictors recommended by this model, and use it to determine values just like we did in the previous example. This model, as we shall discuss later, actually predicts values most efficiently.

2.3. Principal Component Analysis and Linear Regression.

In Principal Component Analysis (PCA), we convert the data into a lower dimension of linearly uncorrelated variables called principal components. Each successive principal components describes less and less of the variance of the data so we pick the top ten components that explain a significant portion of the variability (90 percent or more), to reduce the amount of noise that we deal with. We apply the sliding window approach to PCA and inspect the last 100 days of data for each day we predict a value. For each sliding window, we build a model using the response variable against there other 9 predictors (in this scenario, as their are 10), and then apply the 101 previous days to the 102nd to predict what the price will be.

2.4. Principal Component Analysis and k-Nearest Neighbor Regression.

We once again use the top 10 principal components from Principal Component Analysis, but instead of applying our model coefficients for the past 100 days to a regular multiple linear regression, we instead use the K-Nearest Neighbor Regression, and use our training data to generate predictions.

3. Graphical Interpretation.

We applied the four different methods for predictions on our data. To avoid blowing up the size of this project, we decided to do the analysis on five different currencies and one commodity. From the list, we picked the US Dollar, the Brazilian Real, the Australian Dollar, the Russian Ruble, the Argentine Peso and Oil. The plots for the six different variables are below, with contain the predictions from the different methods as well as the real data.

Plot Legend Titles

- knnPCA - Estimate from Principal Component Analsyis and k-Nearest Neighbor Regression
- lmPCA - Estimate from Principal Component Analysis and Linear Regression
- corLM - Estimate from Pearson Correlation and Linear Regression
- stepLM - Estimate from Forward Stepwise Regression and Linear Regression
- Real Currency- Real Currency or Commodity Price

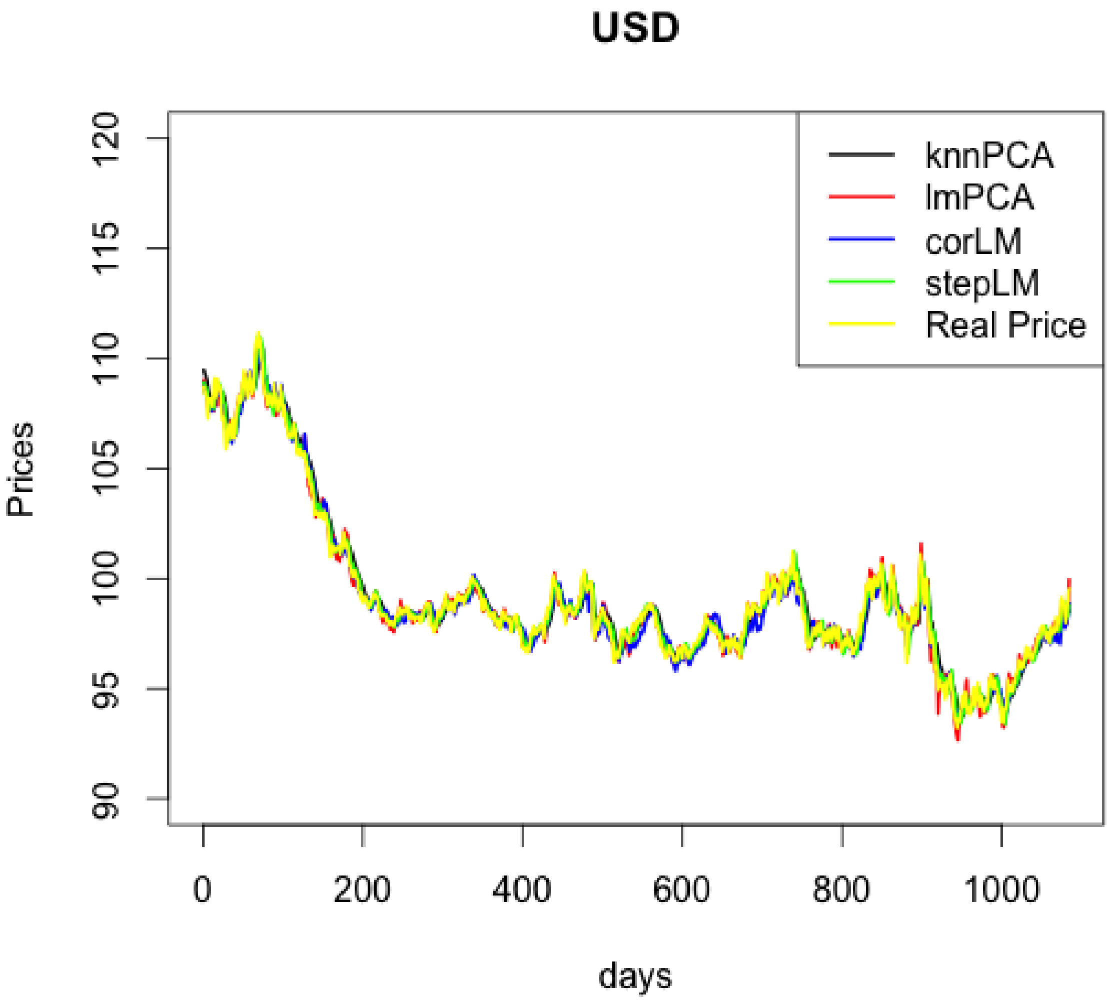

Fig. 3.1: For the United States Dollar, we see that the models are really effective in their prediction. All four methods give us relatively close estimates of the actual REER of the US Dollar, and we can say that this model could be very effective in gauging where the dollar is likely to move next. A close analysis of our error plots indicate that the US Dollar, in fact, is predicted the most effectively among all the variables in consideration.

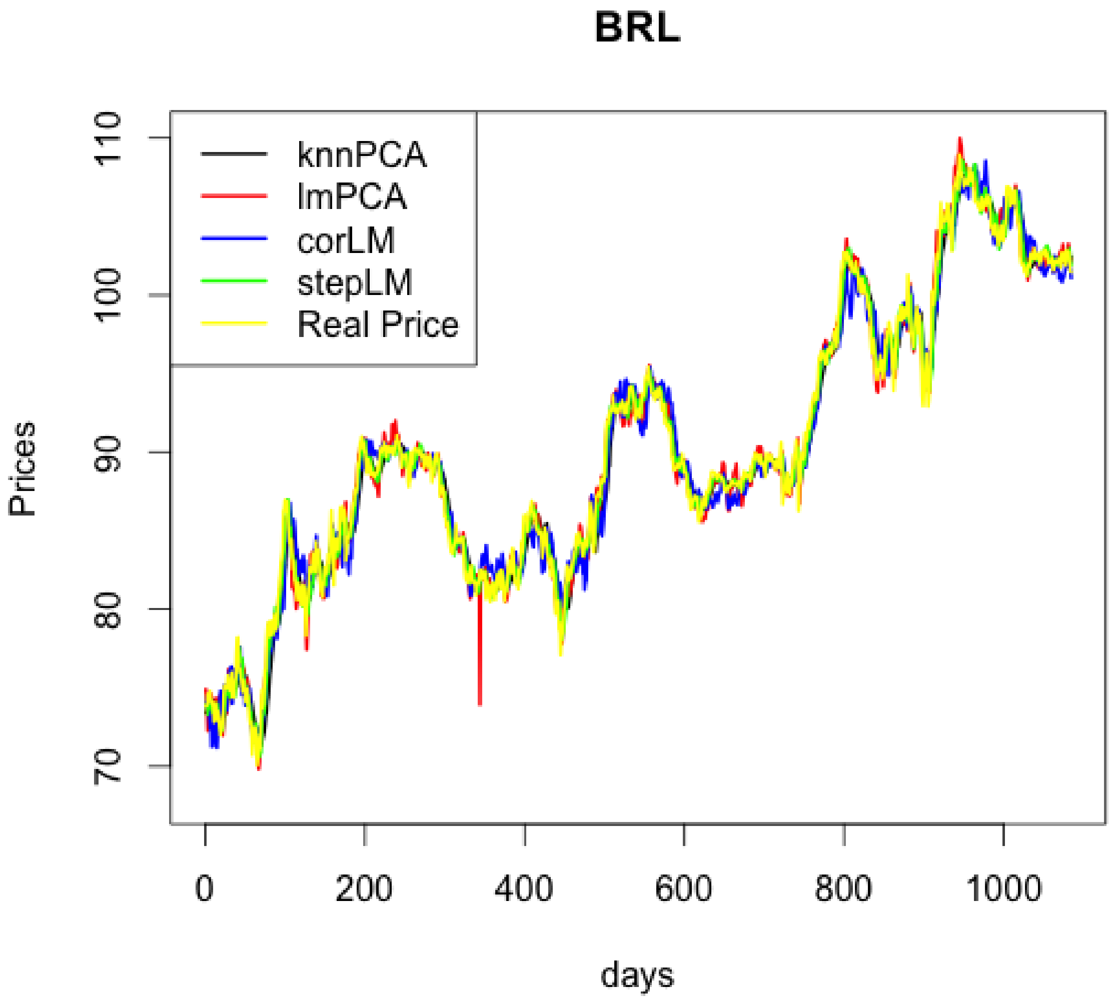

Fig. 3.2: The plots indicate that for the Brazilian Real, the different models are relatively good. They, however, are not as good as the ones for the US Dollar, and this is indicative in the graph for the different models, as well as the error plot, in which the Real has higher error that the USD.

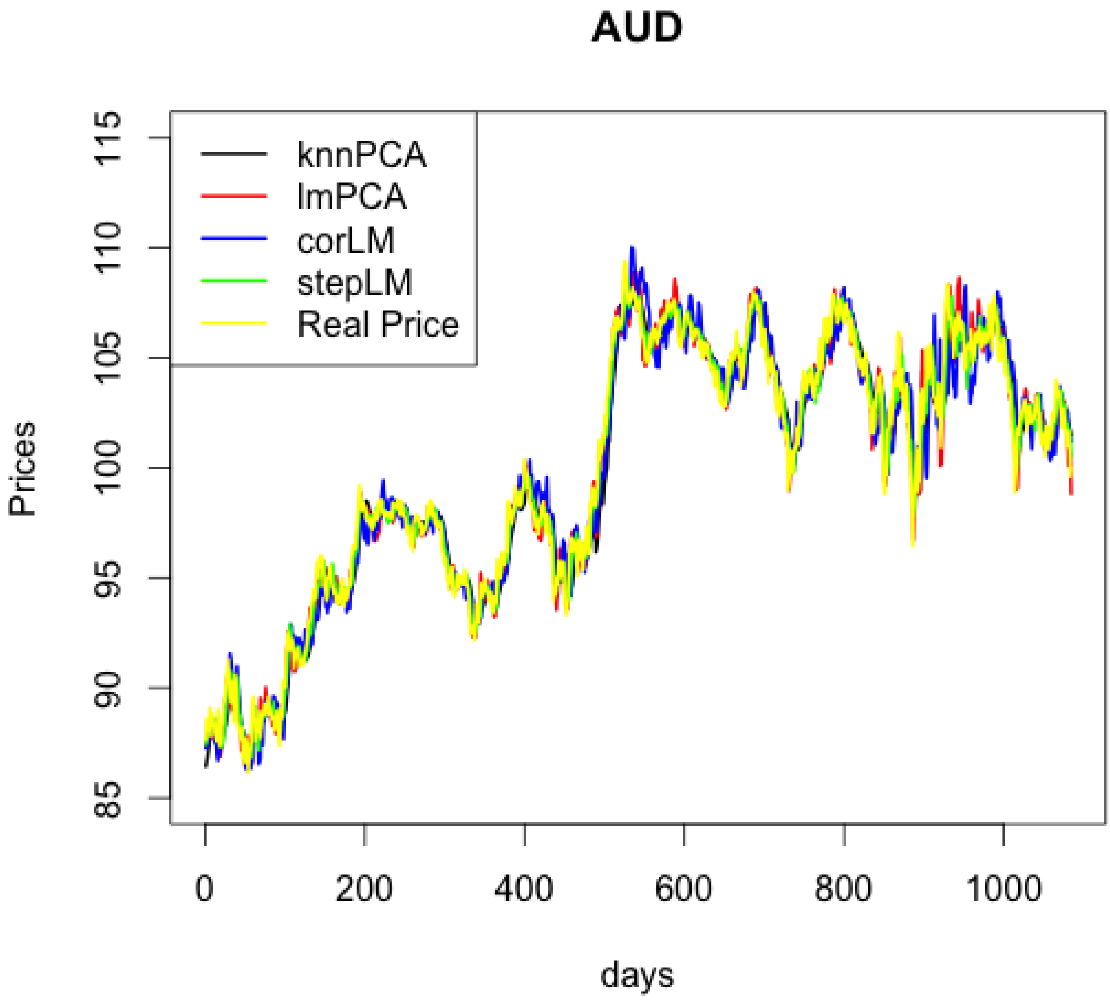

Fig. 3.3: The Aussie Dollar, like the Real, notably has more error in the predictions that we made. After seeing this trend emerge for two currencies that are sensitive to global economic sentiment, especially a variety of commodities, we could start thinking that maybe the models are slightly more erroneous for variables that are more volatile in nature. The error plot reveals the same information.

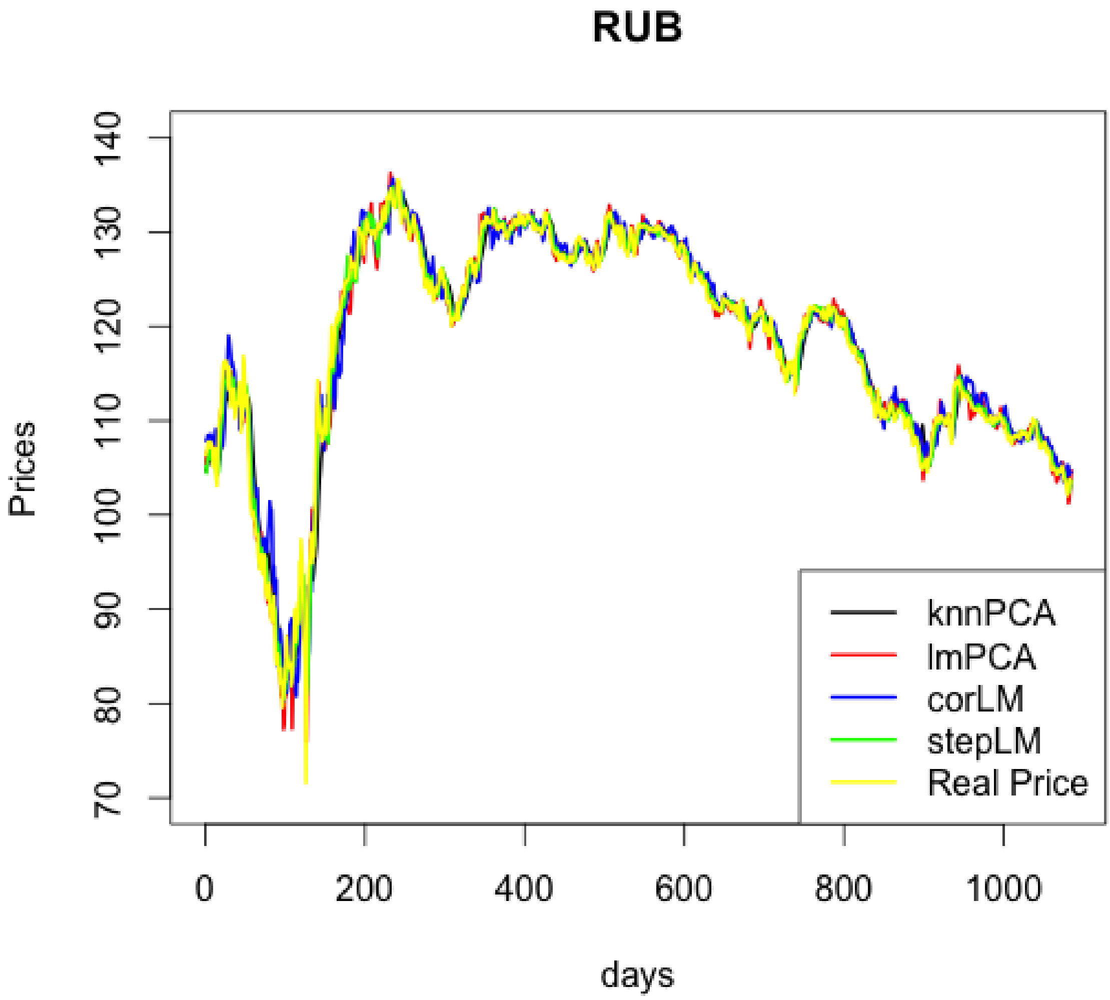

Fig. 3.4: To the naked eye, it seems like the Russian Ruble is predicted well by the different models, but that is not the case when we look at the error data. This backs up our idea well that the higher the volatility of a particular currency, the more likely it is that the error will be more. The Ruble in the last two years has seen a lot of fluctuation because of volatility in oil prices. Oil counts for two-thirds of Russia?s exports. A recent fall in oil prices has affected the Ruble, since a falling demand for the currency has lowered its value. A close reading of Russia?s total exports will show a significant fall, backing up our theory.

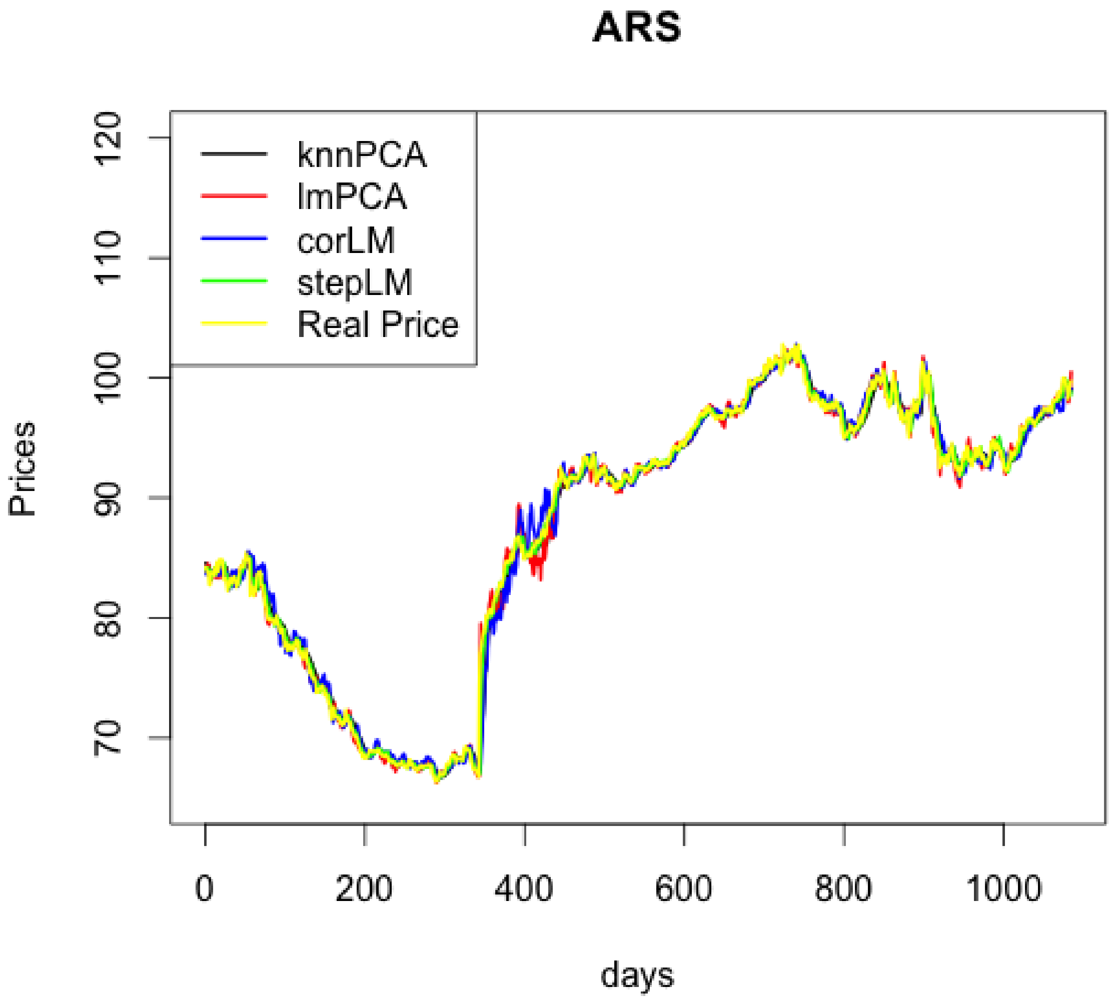

Fig. 3.5: It looks like the Argentine Peso is predicted pretty well by the different models, with the exception of a small window (the volatility might likely have been created during talks of a debt default by the country in that period, which is an external factor and can be hard to account for in a mathematical model). Except that, the model looks pretty good so far, and our mean error plot also reveals relatively lower error for the Argentine currency.

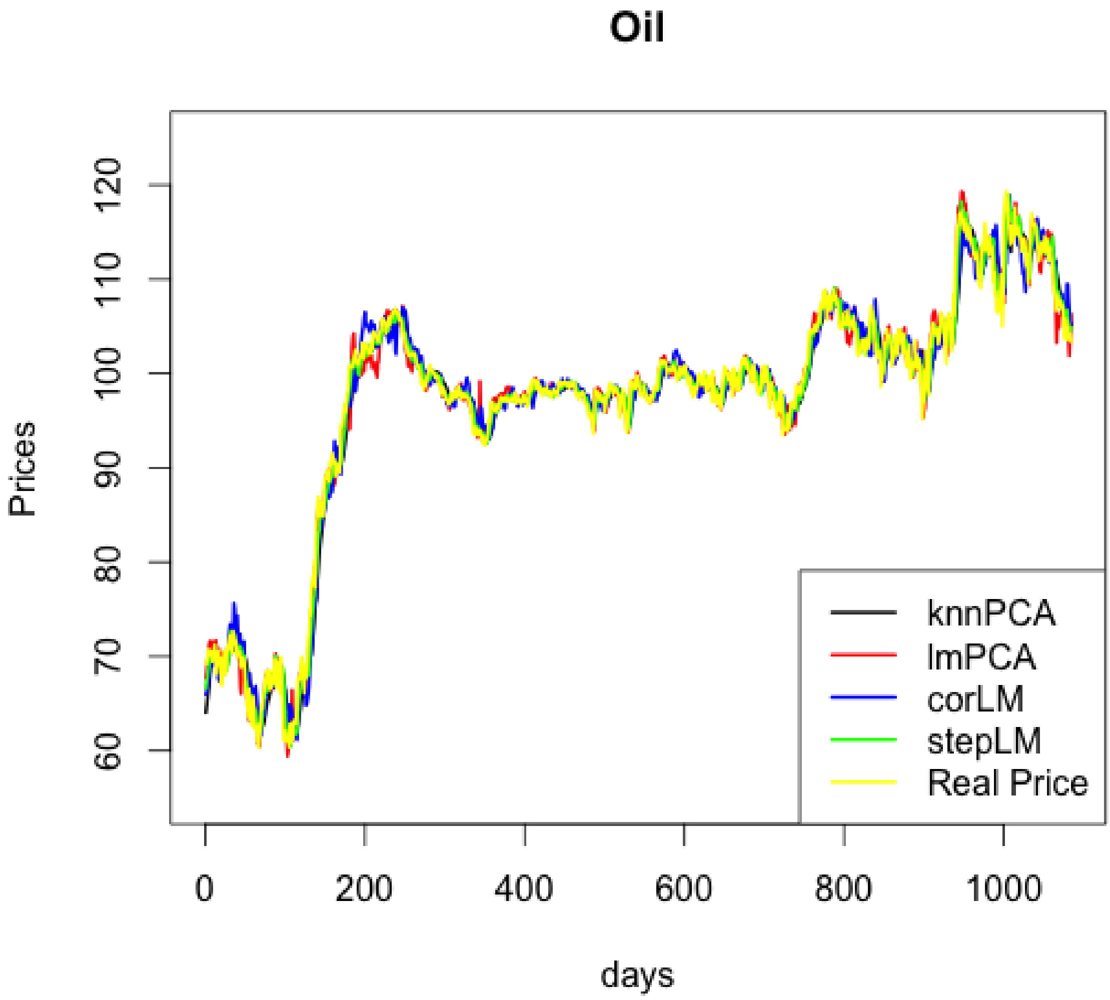

Fig. 3.6: Oil, not surprisingly, is very volatile, and also similar to the Ruble. This is likely because of activity in the oil market over the last two years, as prices have fluctuated due to an increase of production in the United States along with waning global demand and commodity bust. The error plots reveal a mean error of about 0.8, which is similar to Russia?s currency and significantly higher that the US Dollar.

4. Comparison of Model Validity. Once we do our analysis, an important question arises- 'Is our method effective'? We check our results for the average error it produces to make sense To check how good our models are, we run an error analysis on it, comparing it with the actual results. This is simply done by subtracting the estimate from the actual, squaring the difference, and summing it up.

Mean Error Across Currencies and Commodities

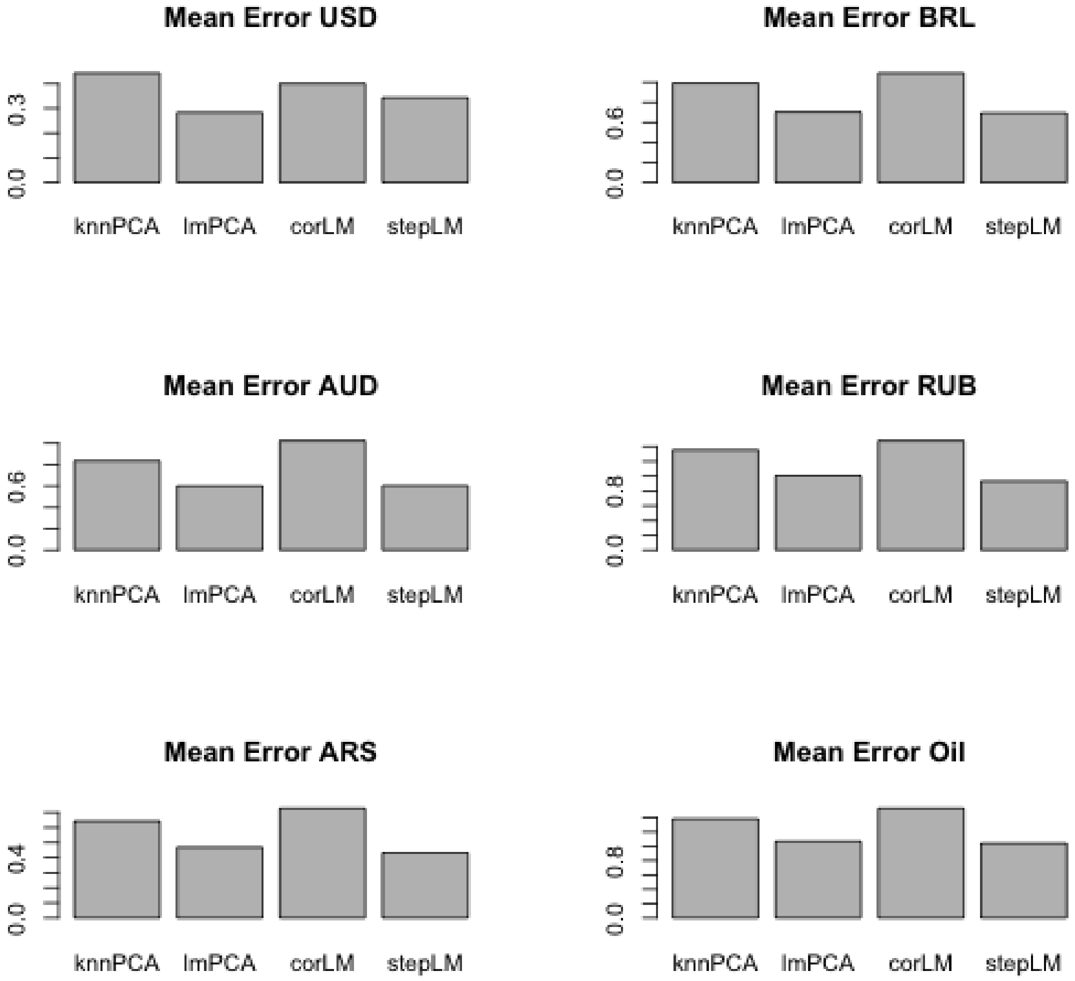

Fig. 4.1: A summary of each currency and commodity displaying the mean error of prediction given method of analysis.

4.1 Errors in Our Prediction On close analysis of our error results, we note that the US Dollar has significantly lower error (around 0.3, for its best model) as compared to the others. The Argentine Peso is next best (around 0.4), followed by the Aussie Dollar (0.6), the Real, the Ruble, and Oil is the highest in the selection we made. This indicates that the error is higher for currencies and commodities that are historically more volatility than the ones that are less. Now, we look at the four models we built individually.

Total Mean Error

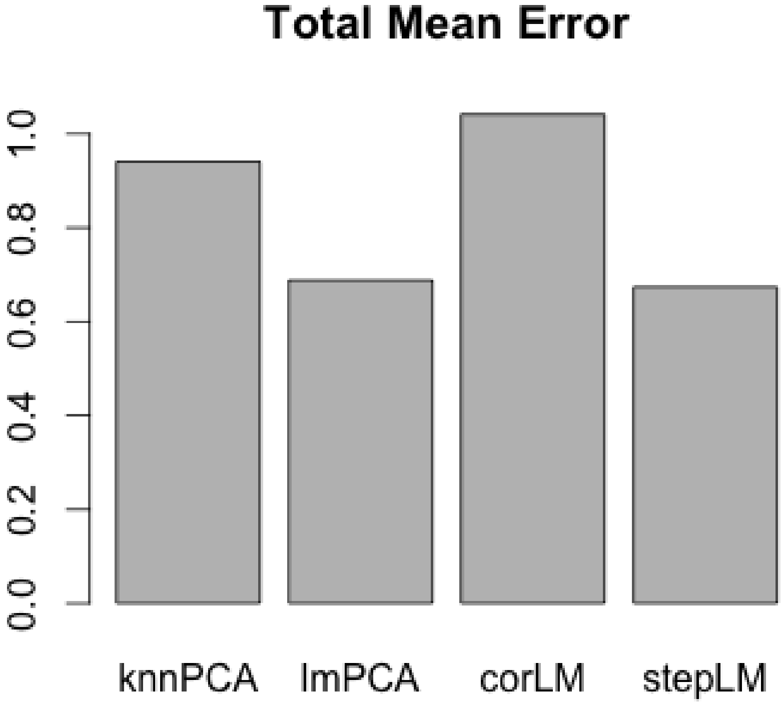

Fig. 4.2: A summary of total error by model across all currencies, showing mean error of prediction given method of analysis.

4.2 Errors in Each Model On average, we note that Principal Component method combined with Linear Regression, and the Stepwise Regression method yield the least erroneous results. We also note that this efficiency in prediction is more pronounced for more volatile items. This tells us that stepwise method (which takes a fair amount of time to compile), is more effective in our goal. Our results make sense - we select predictors that explain our response in the best possible way in the two methods above, which gives us better results.
The combination of K-Nearest Neighbor Regression and PCA yields significantly higher error than the other two method. We think this is because we lost some predictive power due to the nature of KNN Regression which uses the k-nearest neighbors clustering method. Our prediction of highest error is the combination of Pearson Correlation and Linear Modeling. We lose a lot of predictive power in this method because the linear model each time was only built from the top 10 correlated variables. The error with this, though, is that correlation does not tell us a lot about predictive power, and just because variables are correlated, it does not mean that they explain each other well.

5. Conclusion and Next Steps. Our project has given us some exciting results, and hope that testing different models will eventually help us predict prices a lot better. Our models show a certain amount of error in the results, and while it is impossible to completely remove prediction error, there might be ways in which they can be reduced. The forward stepwise method definitely seems effective, and it might be a good starting point for methods being tested in the future.
In terms of next steps, there are several things that can be done with this dataset. One idea that comes to mind is using it to predict prices not for the next day, but for three months or more in advance. That would significantly help companies (that depend on exchange rates and commodities) hedge their bets and neutralize risk. Another idea that comes to mind is using nonlinear methods to predict prices, that may help us build better models.