AFRICAN RECESSIONS

Richard Lawton

First created on 2019-12-24. Updated on 2020-01-01 Submitted in fulfilment of the requirements of HarvardX: PH125.9x Data Science: Capstone course, "Choose your own" assignment.

INTRODUCTION

A recession is defined as a fall in GDP in two successive quarters, and may arise from adverse internal economic conditions, or external factors such as commodity prices. African economies are notoriously susceptible to such external demand fluctuations.

The "African Country Recession Dataset (2000 to 2017)" was created by Chiri in 2019, and hosted on Kaggle:
https://www.kaggle.com/chirin/african-country-recession-dataset-2000-to-2017.

It is a compilation of data from the University of Groningen's Penn World Table Productivity dataset, the Bank of Canada's Commodity Indices and the World Bank's GDP dataset. The goal of the compilation is to answer the question "What factors contribute most to, or are most indicative of, recessions in Africa?"

The dataset follows the ups and downs of 27 African countries over 18 years, recorded as a binary "Recession" or "Not recession" result. (The nation and year in question are not identified.) Each of the 486 observations provides values for 49 contributory variables (described in the appendix to this report). Some of these variables are intrinsic to the nation in question (population, employment, domestic consumption), while others are global (commodity price indices).

The goal of this investigation is twofold. First to build a predictive model of recession. The ability to foresee a recession before it happens is clearly important to a potential investor in the economy - no one plans an investment into a declining economy without very good reason. It is also important to the national government. Recession equates to stagnating or declining living standards, which in an African context can spell real deprivation and hunger, human suffering and political instability. The ability to foresee a recession gives the government at least the potential to take some preventive steps.

The second goal is to gain some insight into the factors that lie behind, or at least correlate with recession. Does the dataset provide us with some useful leading indicators?

In order to achieve these goals a number of steps were taken. Preliminary examination of the dataset revealed some challeges to be addressed: skewed data, imbalanced outcomes and multicollinearity. Careful investigation was required to identify the best way of handling these. Then attention was given to the optimisation criteria to be used: what constitutes a successful predictive model of recession? Using the insight gained from these steps, two machine learning models were optimised - a k-nearest neighbours smoothed model and a random forest model, representing significantly different approaches to the problem. Finally, to examine whether we are able to identify key indicators of recession,

we looked at the importance of the 49 variables as reported by the random forest model, as well as the output of Principal Component Analysis.

METHODOLOGY

Data preparation and overview

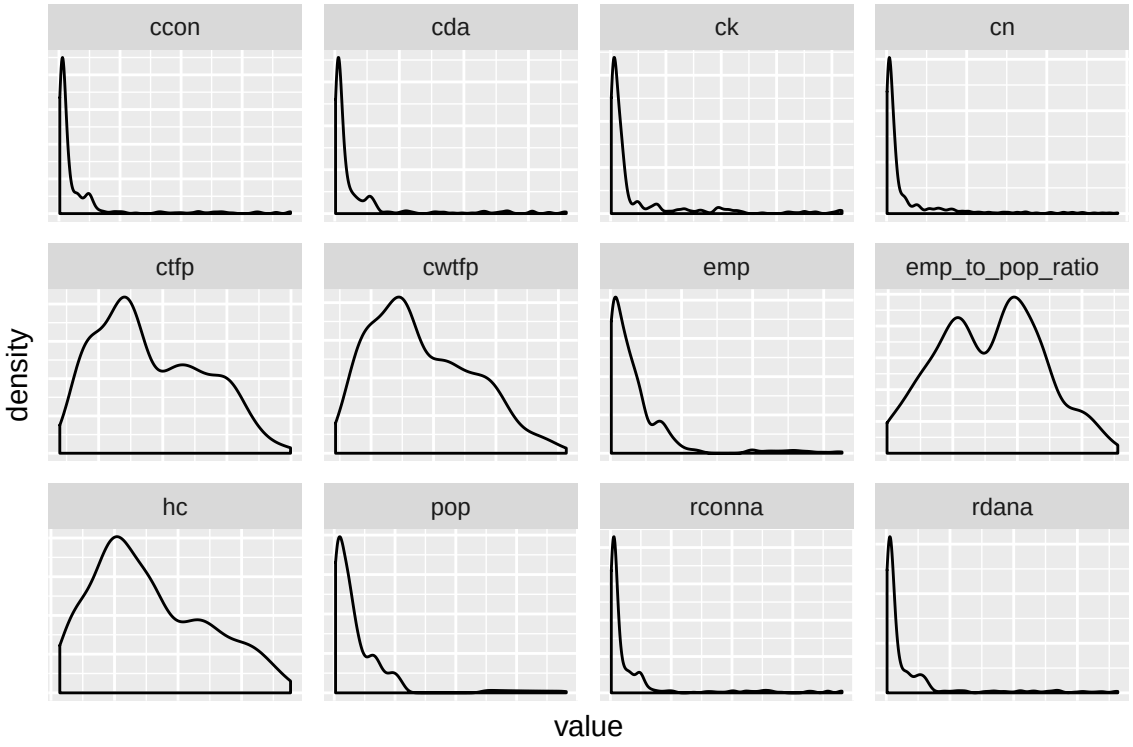
The dataset was downloaded from Kaggle and is made available at https://raw.githubusercontent.com/rtlawton/African-recession/master/africa_recession.csv

There are no missing entries. The result column "growthbucket" was converted to a two factor column, "recession" (positive result) and "growth".

A number of aspects of this dataset become immediately apparent:

1.) There is a significant imbalance of results. The incidence of "recession" in the dataset is 7.82%, while "growth" runs at 92.18%. This will have implications for model stability and predictive power - considerable effort will be invested into creating a model that generates stable, incisive predictions. Central to this issue will be the choice of evaluation metric we use for our model.

2.) Many of the variables have very skewed distributions. For example the first twelve variables are shown here:



This skewness can also affect model stability and predictive power. In order to overcome this a log10 transformation was applied to columns that are skewed according to the

criterion:

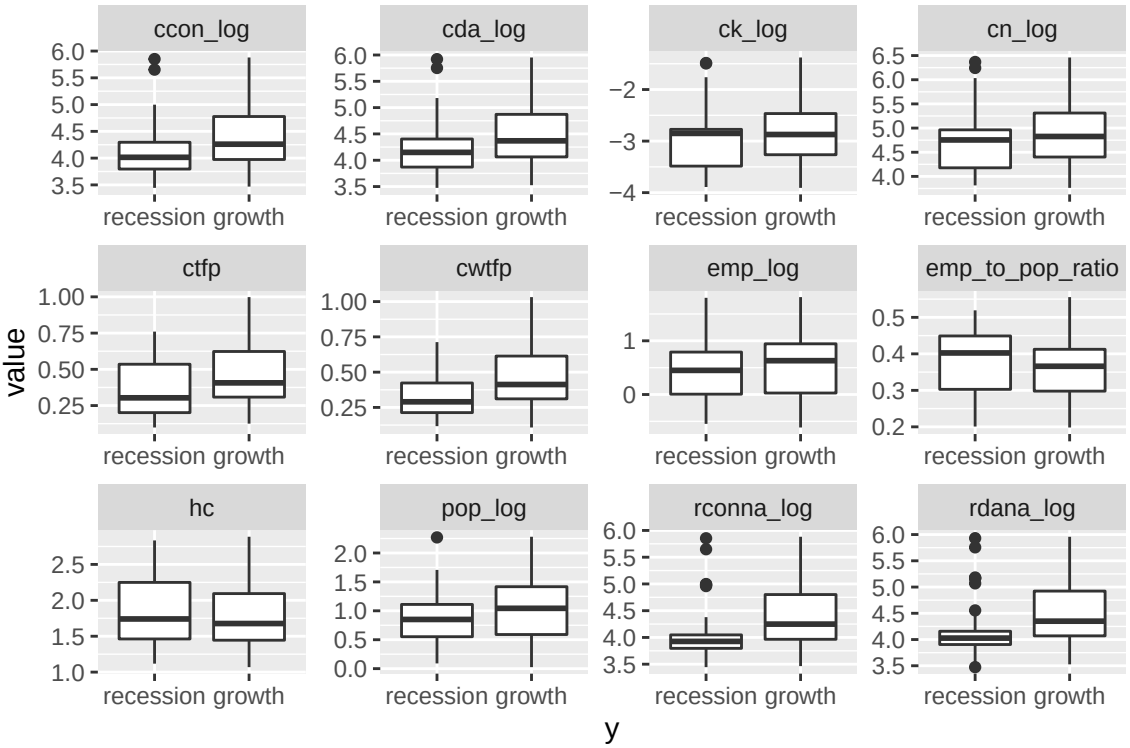
$$max > mean + 4 \times sd$$

or

$$min < mean - 4 \times sd$$

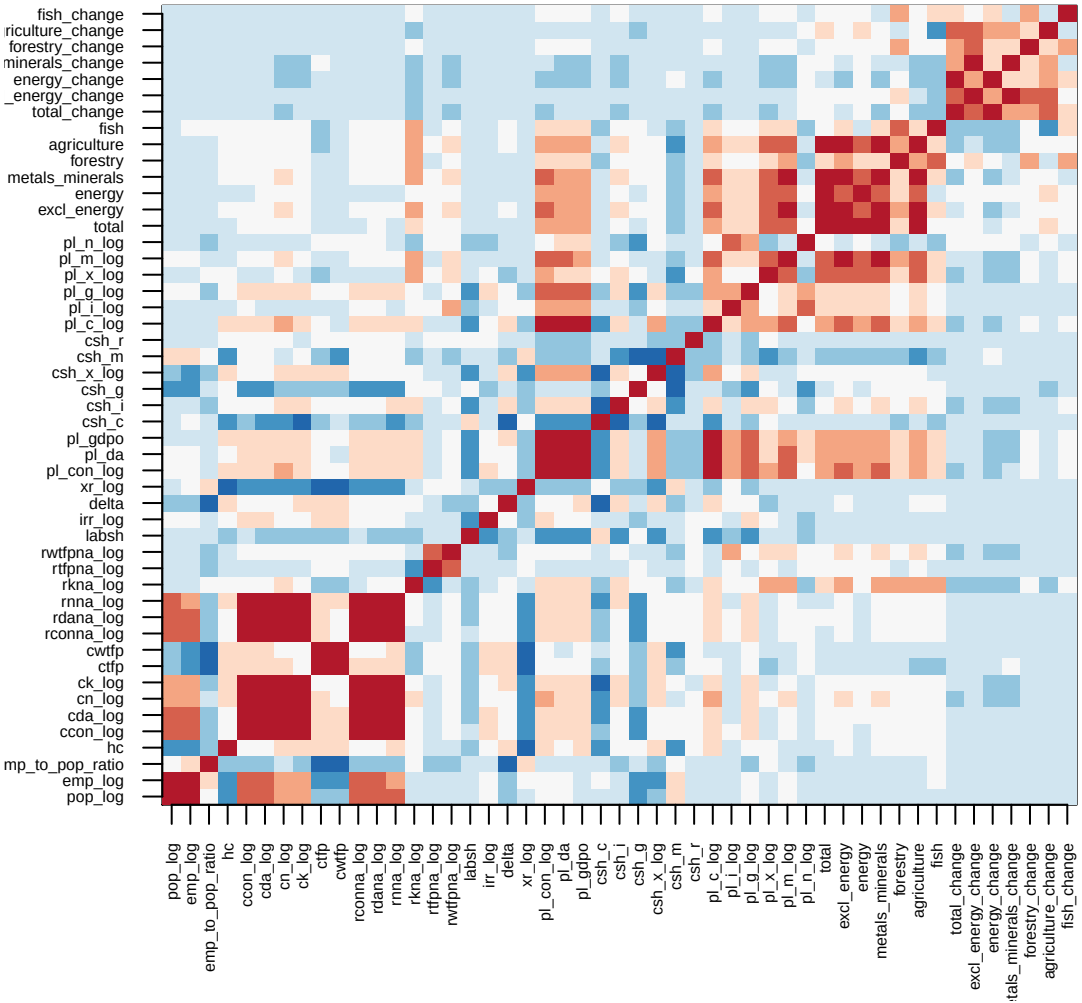
(but not both). Transformed columns were renamed with a "_log" suffix. 22 columns out of 49 were transformed.

3.) None of the variables can function individually as a discriminant between recession and growth, in the sense of having non-overlapping interquartile ranges for the two outcomes. Box-plots for the first 12 variables are shown for example here:



4.) The data columns are heterogeneous, carrying information about diverse matters. As a result the ranges of values recorded vary considerably. For example “ccon” records the real consumption of households and government, measured in million US$ and ranges up to 350,000, while other values such as “emp_to_pop_ration” are fractional. Many of the techniques we will apply will standardise this data to mean = 0 and sd = 1 automatically. Nevertheless we will apply this standardisation ourselves at the outset for the sake of consistency.

5.) There is multicollinearity. For example ccon_log, cda_log, rconna_log and rdana_log are all mutually correlated in excess of 0.986. A full correlation diagram for the variables is shown here:



Will this collinearity affect the models we build? The recognition of collinearity may also offer the opportunity to improve our models through removal of extraneous variables.

After application of the log transform and standardisation, the dataset was divided with stratification into a test set (20%) and a training set (80%). The respective prevalences of recession are 8.16% and 7.73%. No variables were removed from the dataset before this division.

Evaluation metric

The default evaluation metric for a binary classification model is accuracy - the sum of true positives and true negatives divided by all outcomes. For an imbalanced problem this is not a helpful metric, however. With our prevalence of just 8% recessions, we could achieve an accuracy of 92% simply by predicting no recessions at all.

The preferred metric for an imbalanced dataset is the area under the Precision-Recall curve. Precision is the proportion of recessions our model predicts that turn out to be actual recessions, while recall, or sensitivity is the proportion of actual recessions that our model catches. These metrics (precision and recall) highlight false positives and false negatives respectively. Any model will be a trade off between these two metrics - we can guarantee

no false negatives by predicting 100% recession, or no false positives by predicting 0% recession. The area under the curve (AUC) measures how well we have traded one off against the other.

So which is worse - false negative or false positive? To fail to predict a recession that happens carries considerable costs. The external investor is locked into a faltering economy and stands to lose his money. The government has lost economic leverage and opportunity to take preventive action. But the false positive also carries costs - the investor has an opportunity cost of the investment he has not made, and the government has also lost the opportunity to plan its own internal investments.

We would however expect the cost of the false negative to considerably outweigh the cost of false positive. The investor stand to lose 100% of capital through a false negative, while failure to be invested in a given year because of a false positive for recession might only cost him 10-20%. Likewise the government can restrain an overheated economy more easily than stimulate one in recession.

We will therefore also want to monitor recall and precision, to see whether we can build a model that favours recall appropriately.

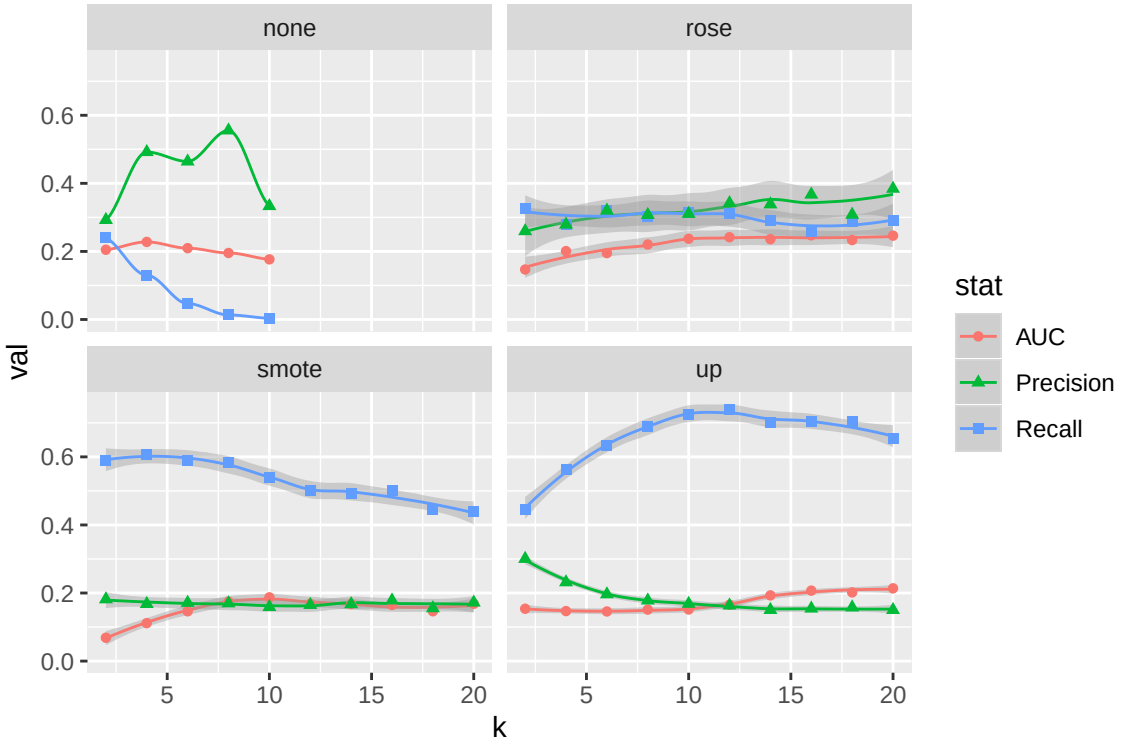
Model stability

The low incidence of the positive class (recession) combined with the relatively small size of the dataset presents some problems for applying machine learning techniques. After we have partitioned out a test set of 20%, we are left with 388 observations of which only 30 will represent the positive class. If we then apply 10-fold cross-validation, each fold will, on average contain only 3 observations of the positive class, with a chance (nearly 5%) that a given fold will contain no positive observations at all. Clearly we will need to have larger rather than smaller folds. For this reason we have chosen to use only 3 folds, but run 10 repeats.

Up sampling techniques

It is accepted practice to improve the performance of machine learning algorithms for imbalanced datasets by rebalancing the positive/negative classes. The first technique to achieve this, called "down-sampling", reduces the number of majority observations. This is unattractive since we do not have a large dataset to begin with. Rather the use of sampling techniques to increase the representation of "recession" outcomes through random repeats of existing observations (up), or random interpolation of new observations (smote or rose) will be preferred.

To test these strategies, we examined a number of training options for a knn model with k=2 to 20. The results are as follows:



The selection of resampling technique clearly has a marked effect on the performance of our model. The baseline model with no resampling demonstrates the need for resampling: the precision is unstable, the recall collapses rapidly with rising k, and the model ceases to function at k=10, presumably because of the lack of any further neighbours.

Rose offers a much more stable picture, but without any noticeable benefit accruing from the knn smoothing technique - adding neighbours doesn't seem to make a great difference. AUC does rise from 0.15 to 0.25 across the range. Recall fluctuates around 0.3 - which means that this model would catch only 30% of all recessions.

Smote offers much better recall - up to 0.6, but this declines with increasing k. AUC rises to a ceiling of 0.2.

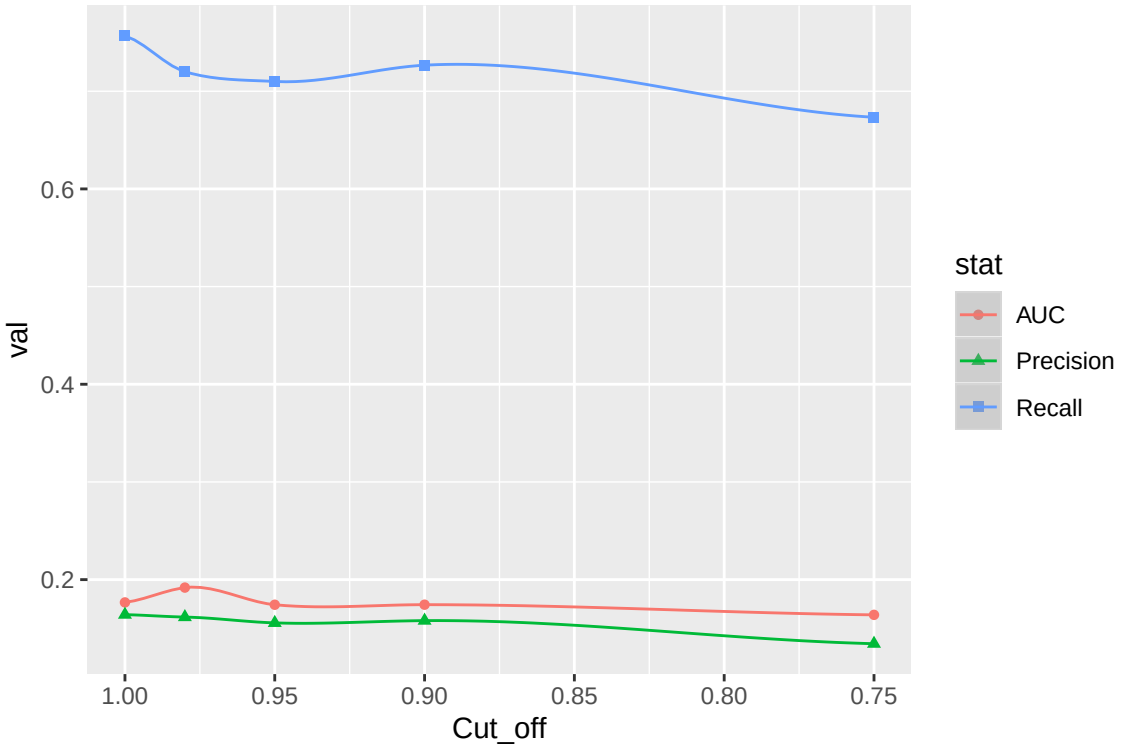
The simple up-sampling technique seems to offer the best option. Recall rises to 0.74 at k=12, with AUC stable at around 0.2. Further model development will use this technique.

Variable reduction

We earlier observed that the data contains multiple collinearities. Can the model be improved by removing some of these? A correlation matrix was constructed and used to identify variables with a high degree of correlation. Some multiple sets of variables exhibited a high degree of mutual correlation. We set successive cut offs for correlation coefficient to remove variables incrementally as follows:

Remove variables:	Retain:
98% cut-off	
cda_log, rdana_log, rconna_log	ccon_log
energy	total
emp_log	pop_log
95% cut-off	
pl_gdpo, pl_da, pl_c_log	pl_con_log
ctfp	cwtfp
cn_log, ck_log	rnna_log
metals_minerals, agriculture	excl_energy
90% cut-off	
energy_change	total_change
rnna_log	ccon_log
75% cut-off	
excl_energy, pl_m_log, pl_x_log	total
excl_energy_change, metals_minerals_change, agriculture_change	total_change
rwtfpna_log	rtfpna_log
fish	forestry
pop_log	ccon_log

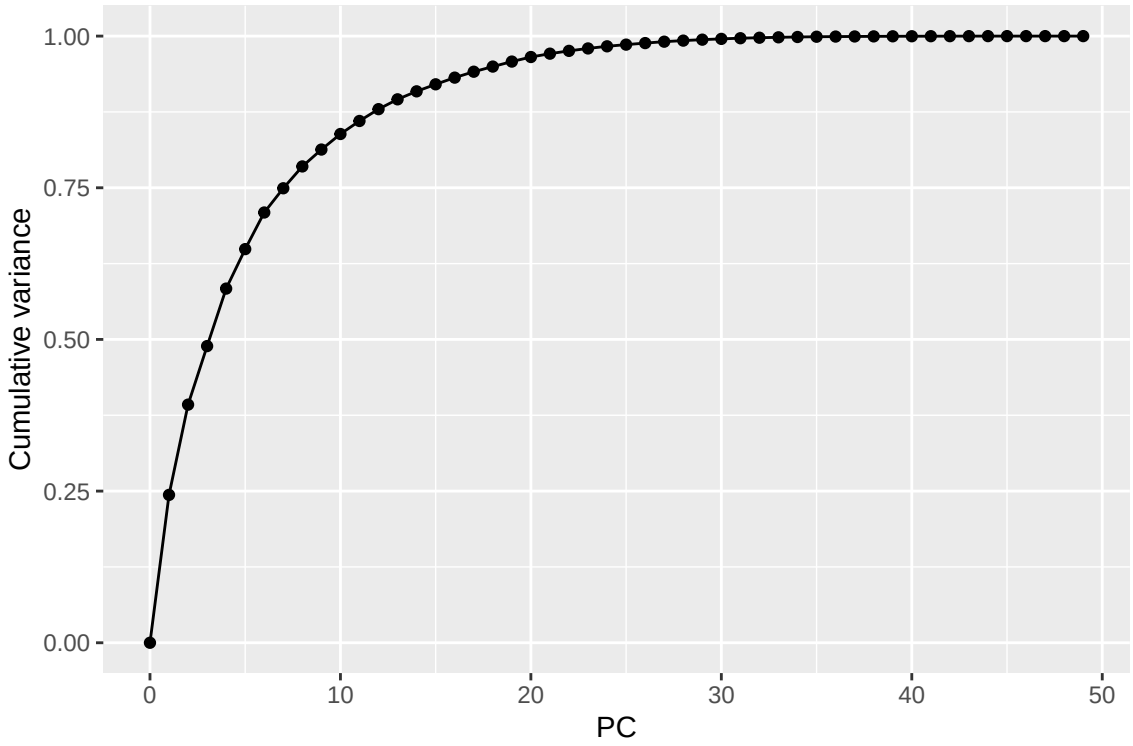
We reran the model for these cut-offs with the following results:



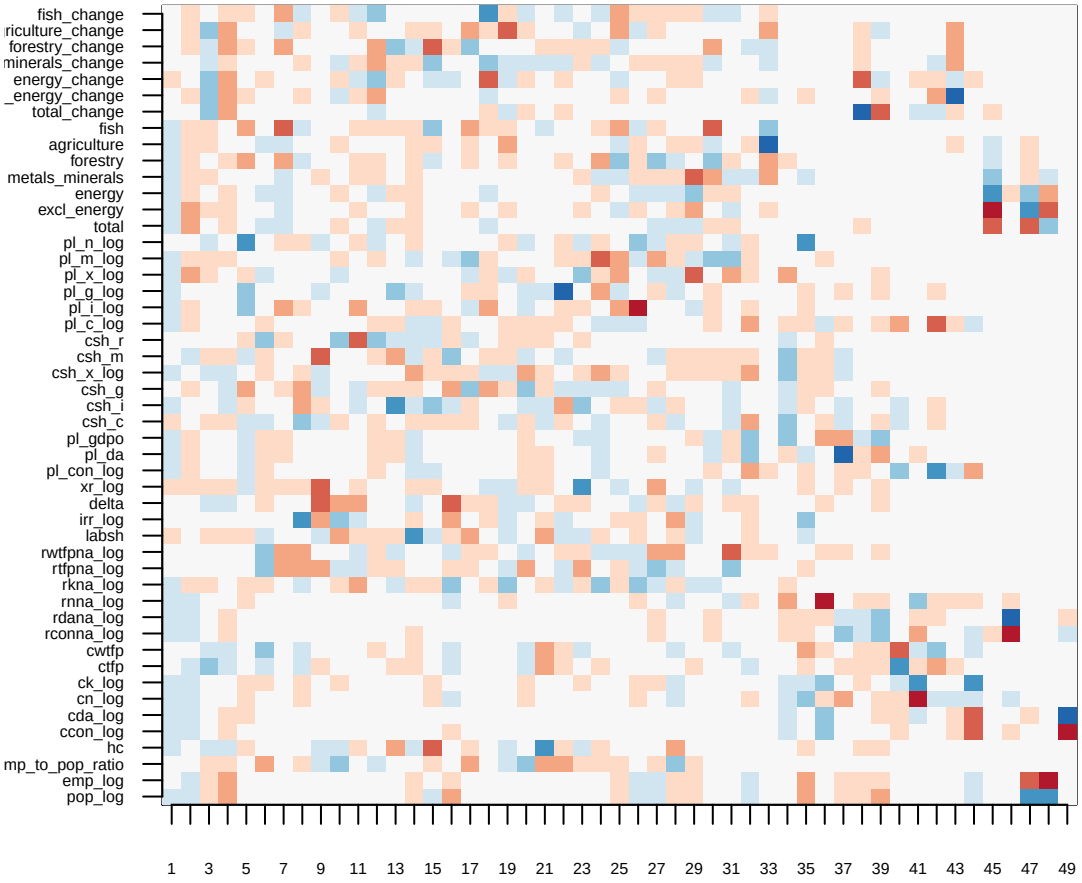
Removing variables has a negative effect on model performance. Recall drops progessively

as variables are removed. Precision and AUC are flat.

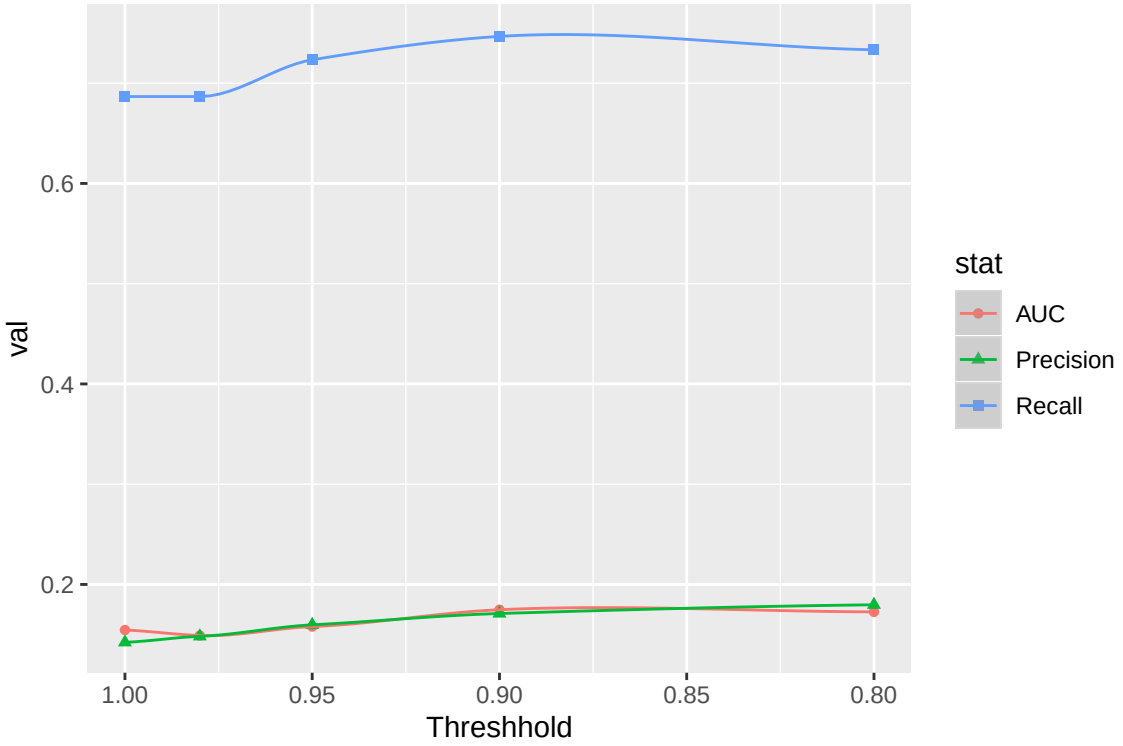
Can we improve the model by applying Principal Component Analysis? Are there variables that do not contribute much information to our model? We ran PCA on our training set with the following results:



This does not help us greatly - to cover 95% of the variance we would need to include 20 out of the 49 PC variables. When we examine the rotation matrix we also see that there is no evidence that some of our input variables feature more strongly in earlier principal components (apart possibly from a combination of commodity price changes in PC2):



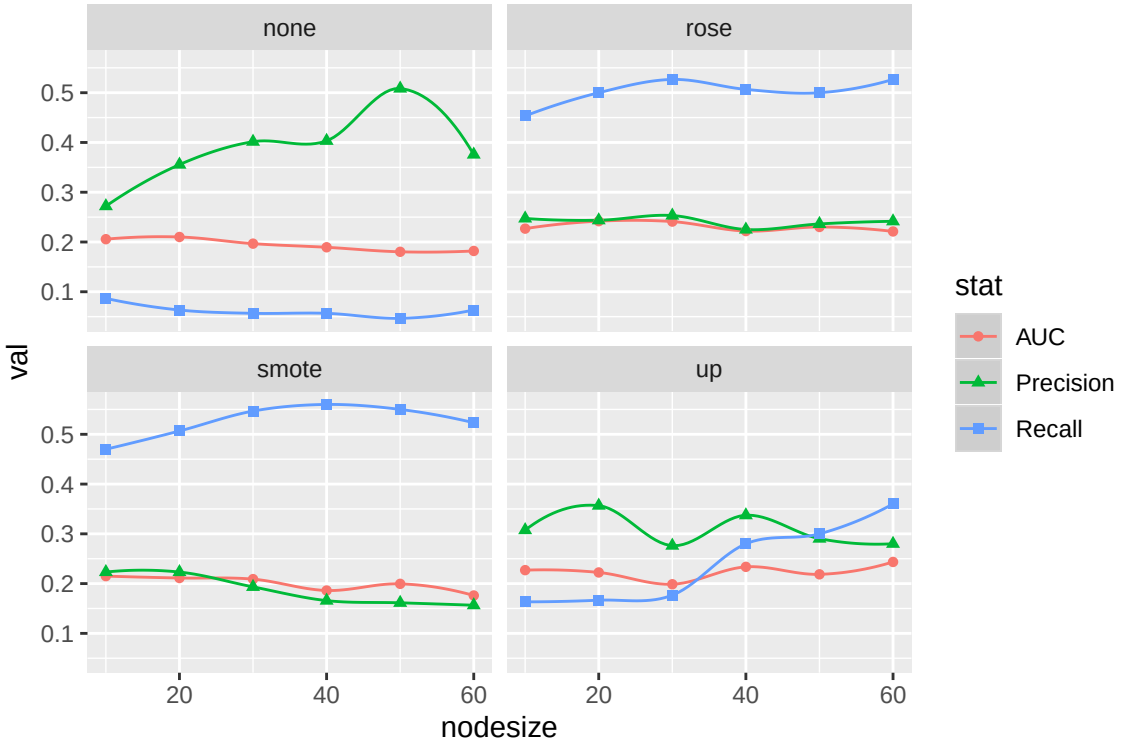
Nevertheless, we tested out the effect on our model, using the PCA pre-processing feature of Caret, using a range of variance cut-offs:



The baseline application of the PCA transformation without removing any components, reduces the recall to 0.687 The recall rises to 0.747 as the cut-off is reduced to 90%, but the benefit gained over our untransformed model is minimal.

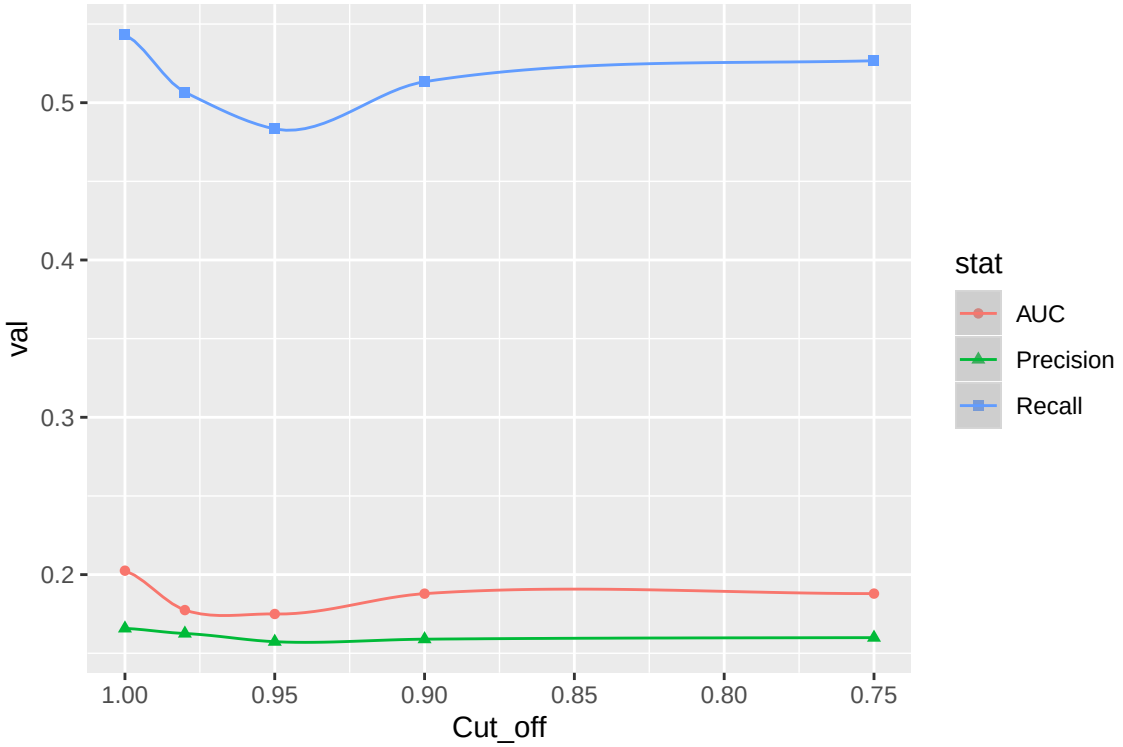
Random Forests

Can we improve our predictions using random forest techniques? Using the Caret package we substituted the "rf" method for "knn" and looked again at the various resampling methods. The training extended over ranges of mtry (3 - 10) and nodesize (10-60). In the event nodesize turned out to be a more useful regulating parameter. The results were as follows:

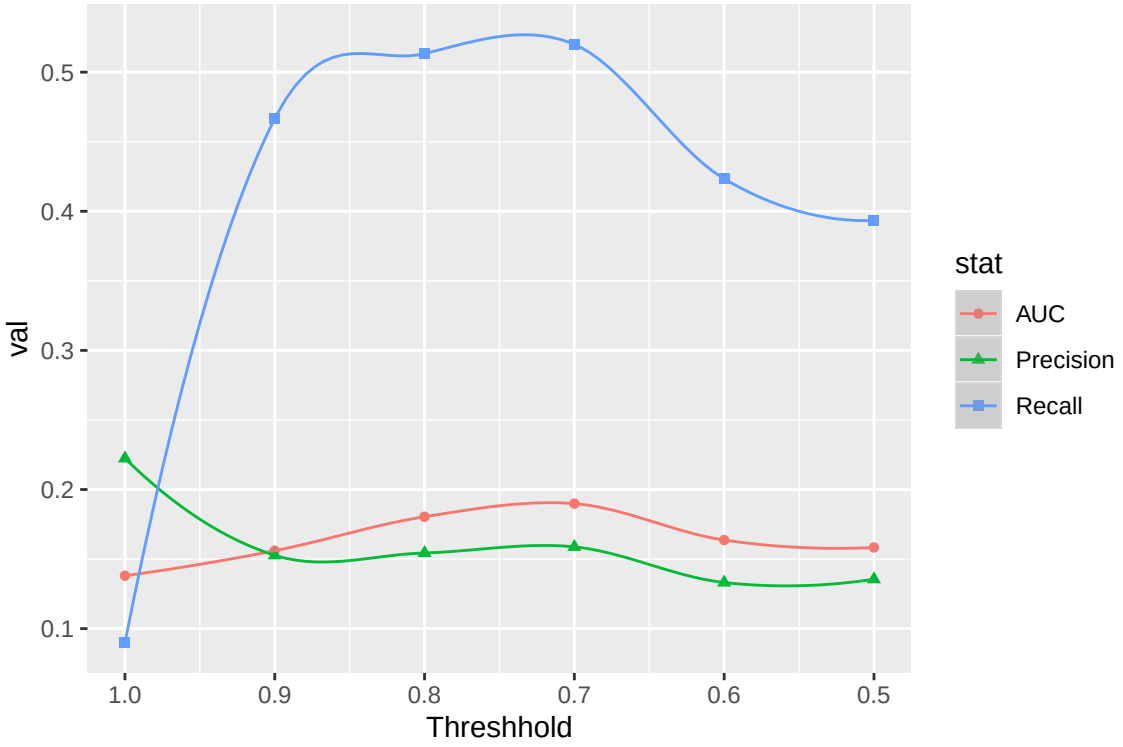


The results are markedly different from the picture we found with knn. In the absence of any rebalancing, the model is more stable than knn, but favours precision strongly over recall, which never rises over 0.1. The simple up-sampling mode also favours precision over recall, but with more variability. A recall of around 0.35 is achieved with a nodesize of 60 (which suggests very simple trees). The best results this time come from rose or smote, which both favour recall strongly over precision. There is actually little to chose between them, but on balance smote does slightly better, with a recall of 0.56 at nodesize = 40. The mtry parameter for this model was 10.

Does dropping variables make any difference to our random forest model? We took our best model parameters from the previous test and applied the model successively to the reduced variable data sets previously generated:



Clearly removing variables does nothing to help our rf model. But will PCA help us with this model? We tried the best rf model with PCA preprocessing, this time with cut offs going down to 50% of variance:



The effect is dramatic. The PCA rotation in itself highly favours precision over recall,

dropping recall from 0.56 to 0.09. As we lower the threshhold so precision falls and recall rises, to a maximum of 0.52 at 70% threshhold. This does not however improve on our native model, which gave us a recall of 0.56.

RESULTS

Predictive power

Our methodology has given us two different models to predict the recession event.

1.) K-nearest neighbours with simple up-sampling and k=12. The model is applied to the dataset, with some variables log10 transformed and the whole set standardised.

2.) A random forest model with smote resampling, mtry=10 and nodesize=40. The same log transformations and standardisation are applied.

How do the models perform on the test data?

The knn model yields the following confusion matrix:

	recession	growth
recession	7	21
growth	1	69

This gives a recall of 0.875, a precision of 0.25 and a AUC value of 0.222.

The rf model yields the following confusion matrix:

	recession	growth
recession	4	19
growth	4	71

This gives a recall of 0.5, a precision of 0.174 and a AUC value of 0.197.

Will an ensemble improve this further? To maximize the recall we will use the rule that a recession is predicted if either knn or rf predict one. This yields the following confusion matrix:

	recession	growth
recession	7	29
growth	1	61

This gives a recall of 0.875, a precision of 0.194 and a AUC value of 0.306.

Recall does not improve further, but sensitivity has worsened. Our best option looks to be the knn model with k=12.

With this model we can say that if the test is passed there is a 25% chance of a recession, but if the test fails, there is only a 1.4% chance of a recession. This is a very discriminating test.

Variable importance

The second goal was to gain some insight as to which variables were more predictive of a recession. We have seen that no variable in isolation provided positive discrimination between the two outcomes. We have seen that there is multicollinearity between various sets of variables, identified above. Beyond this the knn model offers us no insight. The rf model can tell us which variables feature as important, however. The top variables are:

Variable	Importance*
rdana_log	4.78
rconna_log	4.29
energy	2.38
irr_log	2.25
rwtfpna_log	2.15
fish_change	2.14

*Importance is measured as the mean decrease in Gini coefficient associated with that variable

and the bottom variables are:

Variable	Importance
total_change	0.35
pl_gdpo	0.33
forestry_change	0.29
pl_da	0.27
labsh	0.25
metals_minerals_change	0.14

CONCLUSION

We have developed a successful model for predicting the incidence of recession from the data provided, given that a false negative is deemed to be a highly undesirable outcome. Our model will have a failure rate of only 1.4% in detecting recessions. The cost of this high recall is a fairly high incidence of false positives. Our model will misclassify 23.3% growth situations as recession. (By way of reference, the baseline strategy of expecting recession always would achieve 0% false negatives, but 100% false positives.)

One aspect has been hidden in the ambiguous word "predict". We have used it in the sense of guessing the value of unseen (but co-incident) data. In economic terms the word will always have a time value attached. There is little to be gained in telling us we are already in a recession. Much greater utility is to be achieved if we can predict a future recession.

When we examine the relative importance of the variables we find at the very top are "real consumption" and "real domestic absorption", both of which measure the demand level in the domestic economy. We would expect a recession to result in a marked drop in domestic demand, but not necessarily be caused by it. In other words these are co-incident indicators rather than leading indicators. A more interesting study might be to examine

the same range of variables in relation to a recession in the *following* year, but the current data set does not permit that.

APPENDIX - Variable description

Variable	Description
pop	Population (in millions)
emp	Number of persons engaged (in millions)
emp_to_pop_ratio	Ratio of Employed Persons to Total Population
hc	Human capital index, based on years of schooling and returns to education; see Human capital in PWT9.
ccon	Real consumption of households and government, at current PPPs (in mil. 2011US$)
cda	Real domestic absorption, (real consumption plus investment), at current PPPs (in mil. 2011US$)
cn	Capital stock at current PPPs (in mil. 2011US$)
ck	Capital services levels at current PPPs (USA=1)
ctfp	TFP level at current PPPs (USA=1)
cwtfp	Welfare-relevant TFP levels at current PPPs (USA=1)
rconna	Real consumption at constant 2011 national prices (in mil. 2011US$)
rdana	Real domestic absorption at constant 2011 national prices (in mil. 2011US$)
rnna	Capital stock at constant 2011 national prices (in mil. 2011US$)
rkna	Capital services at constant 2011 national prices (2011=1)
rtfpna	TFP at constant national prices (2011=1)
rwtfpna	Welfare-relevant TFP at constant national prices (2011=1)
labsh	Share of labour compensation in GDP at current national prices
irr	Real internal rate of return
delta	Average depreciation rate of the capital stock
xr	Exchange rate, national currency/USD (market+estimated)
pl_con	Price level of CCON (PPP/XR), price level of USA GDPo in 2011=1
pl_da	Price level of CDA (PPP/XR), price level of USA GDPo in 2011=1
pl_gdpo	Price level of CGDPo (PPP/XR), price level of USA GDPo in 2011=1
csh_c	Share of household consumption at current PPPs
csh_i	Share of gross capital formation at current PPPs
csh_g	Share of government consumption at current PPPs
csh_x	Share of merchandise exports at current PPPs
csh_m	Share of merchandise imports at current PPPs
csh_r	Share of residual trade and GDP statistical discrepancy at current PPPs
pl_c	Price level of household consumption, price level of USA GDPo in 2011=1
pl_i	Price level of capital formation, price level of USA GDPo in 2011=1
pl_g	Price level of government consumption, price level of USA GDPo in 2011=1
pl_x	Price level of exports, price level of USA GDPo in 2011=1
pl_m	Price level of imports, price level of USA GDPo in 2011=1
pl_n	Price level of the capital stock, price level of USA in 2011=1
total	Annual Bank of Canada commodity price index - Total
excl_energy	Annual Bank of Canada commodity price index - Excluding Energy
energy	Annual Bank of Canada commodity price index - Energy
metals_minerals	Annual Bank of Canada commodity price index - Metals and Minerals
forestry	Annual Bank of Canada commodity price index - Forestry
agriculture	Annual Bank of Canada commodity price index - Agriculture
fish	Annual Bank of Canada commodity price index - Fish
total_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Total
excl_energy_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Excluding Energy
energy_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Energy
metals_minerals_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Metals and Minerals
forestry_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Forestry
agriculture_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Agriculture
fish_change	Year-on-Year Percentage Change Annual Bank of Canada commodity price index - Fish
growthbucket	"1" = Recession; "0" = No_Recession