Piergagnon Coulibaly

pcoul2@uis.edu

Introduction to Machine Learning

Abstract

Using Machine Learning Models to predict when a country is in recession

Final Project Report

Predicting Recession of African Countries

Contents

[**Abstract** 2](#_Toc70758000)

[***Problem Definition* and *Project Goals*** 2](#_Toc70758001)

[***Related Work*** 3](#_Toc70758002)

[***Data Exploration and Preprocessing*** 4](#_Toc70758003)

[***Data Analysis and* Experimental *Results*** 16](#_Toc70758004)

[**1.** **Classification Models** 16](#_Toc70758005)

[**a.** **KNN without PCA** 16](#_Toc70758006)

[**b.** **KNN With PCA** 17](#_Toc70758007)

[**c.** **Naïve Bayes Without PCA** 17](#_Toc70758008)

[**d.** **Naïve Bayes With PCA** 17](#_Toc70758009)

[**2.** **Regularization Models** 17](#_Toc70758010)

[**a.** **Lasso Model** 18](#_Toc70758011)

[**b.** **Ridge Model** 18](#_Toc70758012)

[**c.** **Elastic Net Model** 18](#_Toc70758013)

[**3.** **Ensemble Models** 19](#_Toc70758014)

[**a.** **Random Forest** 19](#_Toc70758015)

[**b.** **Gradient Boost** 19](#_Toc70758016)

[**c.** **Ada Boost** 19](#_Toc70758017)

[**d.** **Tree Bag** 19](#_Toc70758018)

[**4.** **Neural Network** 20](#_Toc70758019)

[**Models Comparison** 20](#_Toc70758020)

[**Conclusion** 21](#_Toc70758021)

[**References** 23](#_Toc70758022)

## **Abstract**

**The project was designed in an attempt to tackle the problem of recession in the African continent. Specifically, whether an upcoming recession could be predicted based on a set of economic features. A good number of models were built and trained on a set of training data. The models’ predictions were also compared against a set of unseen data known as test data. The results of the experiments revealed that class imbalance can seriously hamper the performance of even the best machine learning models. Therefore, a good enough compromise in terms of model performance had to be chosen.**

## **Problem Definition and Project Goals**

**The main motivation behind this study is for potential business investments. A recession is viewed as a decline in economic activity that generally leads to negative gross domestic product (GDP), falling retail sales and rising levels of unemployment**[[1]](#footnote-1)**. The other way to look at it from a business standpoint is that recession can lead to opportunities. These opportunities can range from slashing materials and supply chain costs to exploiting markets left untouched by competitor**[[2]](#footnote-2)**. In other words, recession is a potential investment opportunity if acted upon early on. On the other hand, a recession can badly hurt consumer confidence in a country’s economy. This point is important because it will determine the cost of the positive and negative rates in our machine learning models. The dataset used for this analysis was obtained from Kaggle. It blends the University of Groningen's Penn World Table Productivity dataset, the Bank of Canada's Commodity Indices and the World Bank's GDP dataset**[[3]](#footnote-3)**. The features in the dataset were collected from the year 2000 to 2017. They are objective economic indices that do not include any column identifying a particular country. Many of these features do not share the same numeric scale as a result of being economic indices. Therefore, it goes that a great deal of numeric scaling will be necessary. The scaled data will then be used to train various machine learning models with the intent to predict whether a country is in recession or not.**

**I will set the no recession cases as the positive class. It would be best to have a high false negative, ‘yes recession’ misclassified as a ‘no recession’, rather than a high false positive, “no recession” misclassified as a “yes recession”. Therefore, the cost of misclassifying a country in recession as not in recession is better than misclassifying a country not in recession as in recession. This is mainly because it would be costly financially for businesses in the long run to prepare for a recession when there is none. For consumers, the thought of a recession would damage confidence in terms of spending and overall hurt the economy.**

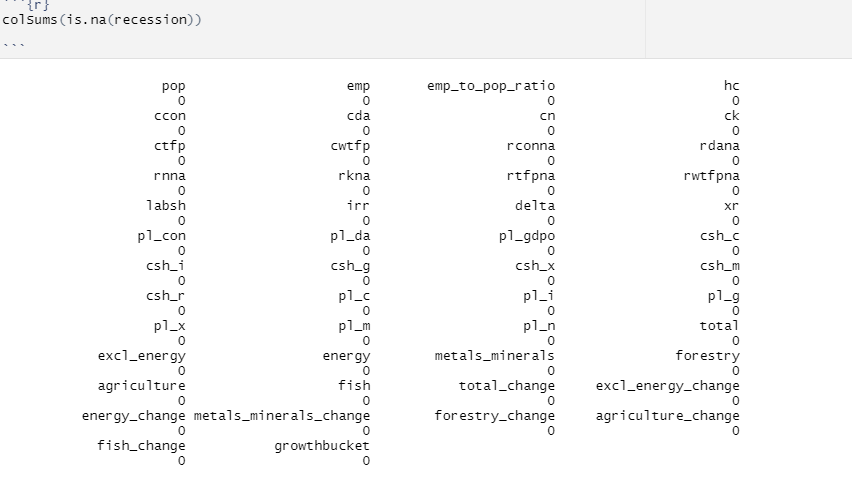
## **Related Work**

**There is one study on Kaggle that specifically addresses the problem formulated in this study**[[4]](#footnote-4)**. The approach of that study was oriented toward an extraction of features to better understand the relationship between the outcome variable and the other features. The author used the ggplot techniques mainly for data exploration and feature extraction. Once a criteria was determined for feature separation, the author then devised a formula to keep some features for the model. Unlike my approach, there were no specific separation of data between train and test. Throughout the study, the author briefly mentioned the class imbalance issue that I believe is critical in the way similar studies should be conducted. However, the same SMOTE technique was used to deal with the imbalance.**

**The author cautiously approached her model with the acknowledgement of her limitations in understanding the meaning of some of the economic indicators inside the data set. It appears that limitation is a driving factor in the low confidence shown in the final model. Albeit a good point, I argue that a focus on the data alone should be sufficient in elaborating a model and measuring its performance without having above average expertise in the field of economics. Another difference between the study’s model and my model is the author used accuracy as a barometer of performance. The Kappa statistics is a better gauge of performance especially for data where the outcome feature is imbalanced**[[5]](#footnote-5)**. With that in mind, I used some additional metrics, false positive/negative rates and Area Under Curve, to give my models a more complete and accurate estimation. Similarly, the author of the study did acknowledge that any model tested on this data set would not be perfect or high performing.**

## **Data Exploration and Preprocessing**

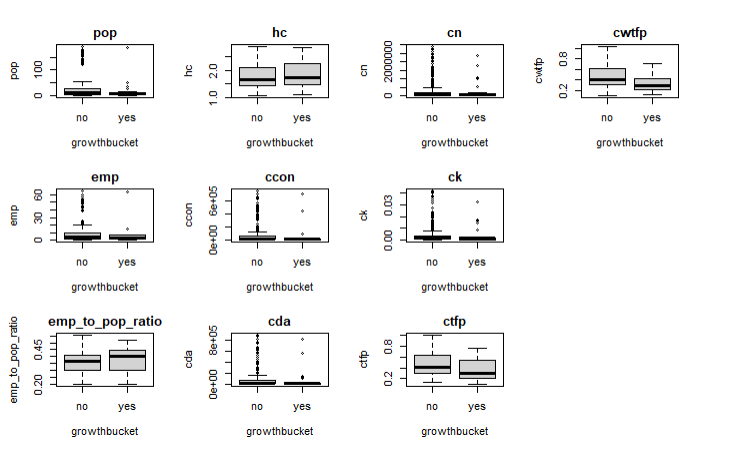
**The data set called recession has 486 observations with 50 variables. All variables in this set are numeric except growthbucket, which will be the outcome variable. That variable is a factor of two levels. A ‘0’ factor indicates no recession while a ‘1’ indicates a recession. For the purpose of simplicity in building the models, I decided to change the factors to a ‘no’ which corresponds to no recession and a ‘yes’ meaning there is a recession. An overview of the data set seems to point that luckily there are no missing values.**

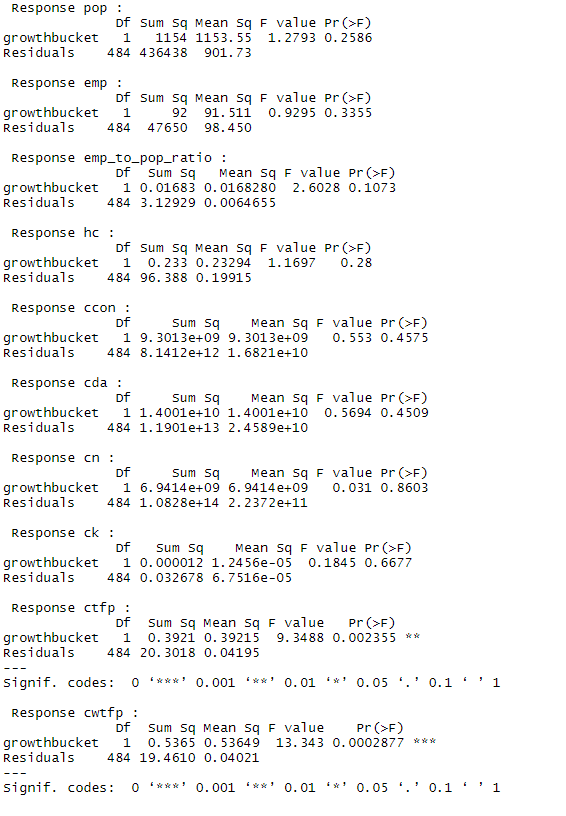


**A look at the outcome variable revealed a highly imbalance class. There are a total of 92.1% responses with ‘no’ values and only 7.8% of ‘yes’ values. This imbalance would determine the approach taken in selecting features to build the models.**

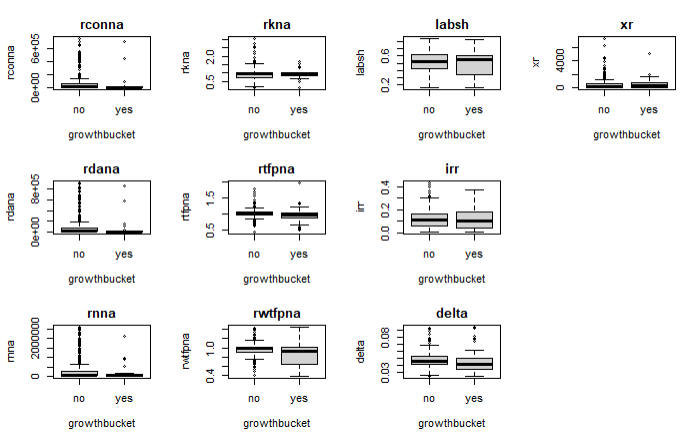
**The next step in the exploration phase was to examine any potential relationship between the outcome feature and all other variables. To accomplish this task, side by side box plots were used in addition to multiple ANOVA tests (MANOVA). Then, the individual one way ANOVA results were extracted using the summary function. The continuous variables were separated into 5 groups.**

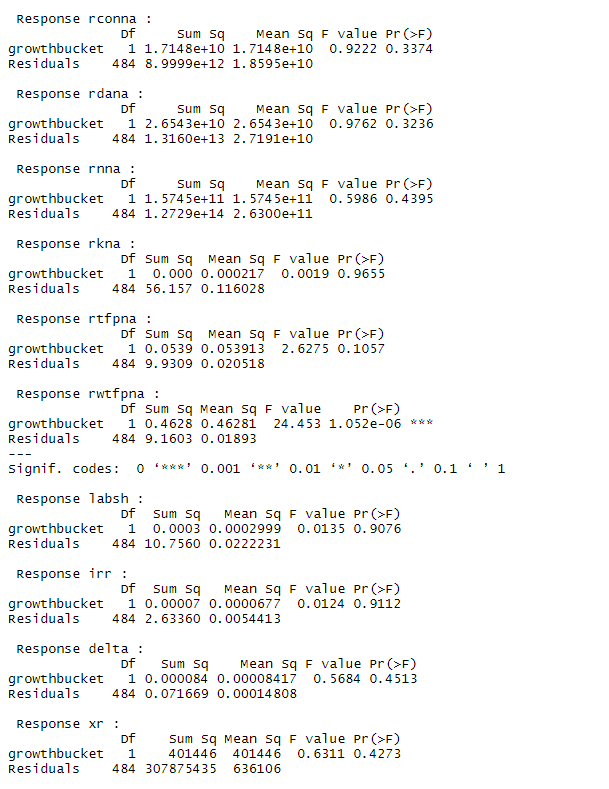
**Group 1 plots and statistical tests: Only ctfp and cwtfp have a confirmed significant difference.**



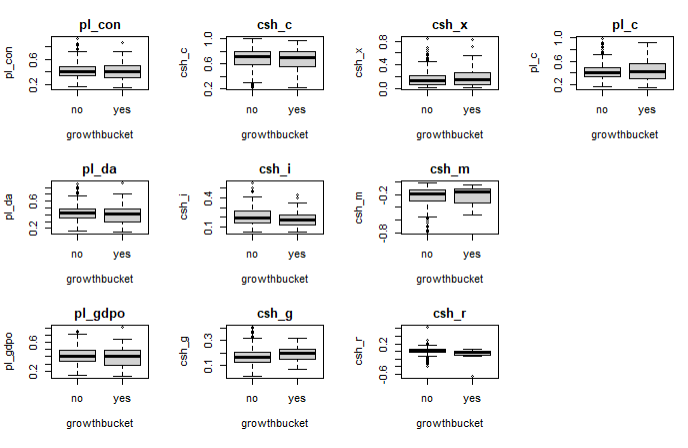


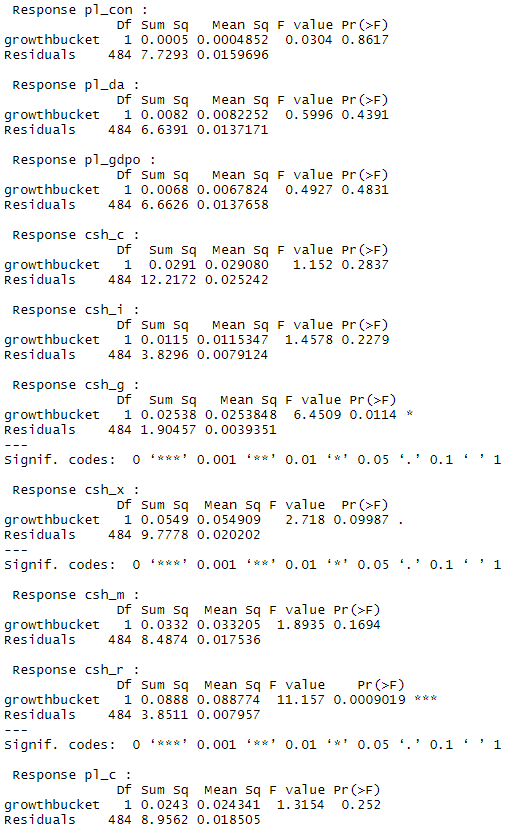
**Group 2 plots and statistical tests: only rwtfpna has a statistical difference with growthbucket and would be kept as predictor.**



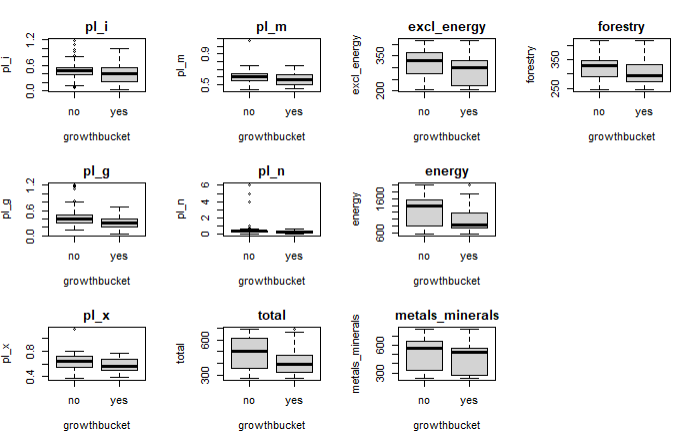


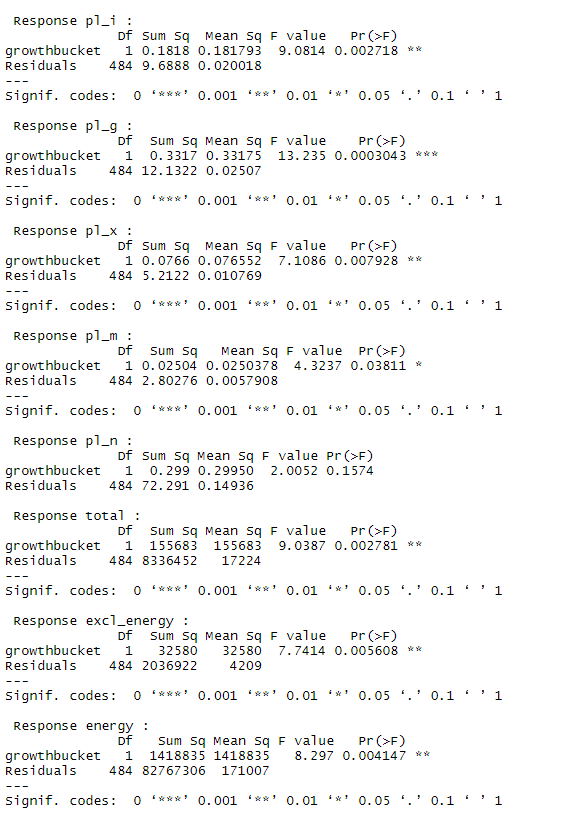
**Group 3 plots and statistical tests: Only csh\_g and csh\_r would be kept in this group.**

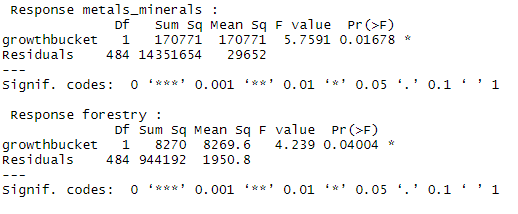




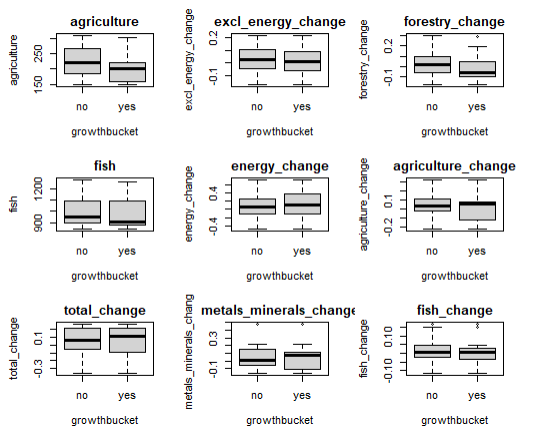
**Group 4 plots and statistical tests: All features in this group would be kept based on the observations of the box plots and the results of the MANOVA test.**

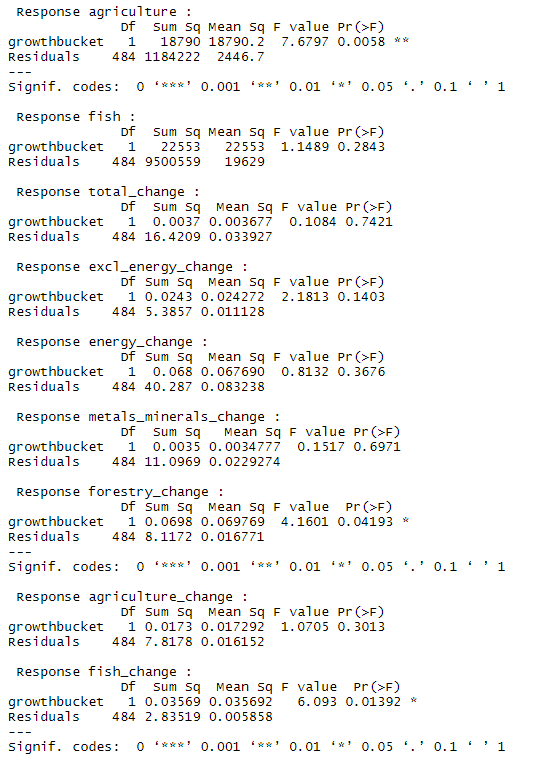






**Group 5 plots and statistical tests: Only agriculture, forestry\_change and fish\_change have a statistical difference with growthbucket based on the statistical tests.**





**These observations and statistical tests reveal that only 18 variables have a relationship with the outcome variable. That would leave out 31 variables. Given that the data set only has 486 observations, that would be a lot of data and potential valuable information to discard. A potential explanation for this lack of relationship could be the class imbalance of the outcome variable. Therefore, it would be wise to keep all variables.**

**In order to mitigate the class imbalance in the outcome variable, the SMOTE method of creating synthetic data was adopted. That method is available in the caret package and was included within the train control steps of creating the folds for the models. Another preprocessing step was scaling all numeric features. The Scale method in caret was used to accomplish that task. These two steps were applied only to the train set after splitting the original data set into 80% training set and 20% testing set.**

## **Data Analysis and Experimental Results**

**A total of 12 models were built for this study from 4 main model groups: Simple classification, regularization, ensemble and neural network. The classification, regularization and ensemble models were trained using the train function from the caret package. The predict and confusionMatrix functions were applied on the models to predict the outcome variable. Model performance was measured using the Kappa statistics, Area Under Curve (AUC) and False Positive/False Negative Rates.**

### **Classification Models**

**Two of the classification models were built with Principal Component Analysis (PCA) as a preprocessing step in the train function. The motivation behind using PCA was to take advantage of dimensionality reduction. Therefore, one model was used with PCA and the other without PCA.**

### **KNN without PCA**

**The model hyper parameters were auto tuned by caret with the k set to 7 for the final model. The Kappa statistics was 0.17 and AUC of 0.83. This means the model was deemed excellent at distinguishing between true positive and false positive while having a slight agreement between predicted and actual values. The false positive rate (FPR) was 0.01.**

### **KNN With PCA**

**When the PCA parameter is added, the model’s kappa remained the same at 0.17 and the AUC slightly increased to 0.86. Regardless of these changes, the model would still be classified has having slight agreement in terms of kappa and excellent/good in terms of AUC. The FPR did not change.**

### **Naïve Bayes Without PCA**

**In this model, the auto tuned parameters were o for laplace, TRUE for usekernel and 1 for adjust. The Kappa statistics of 0.25 points to a model with fair agreement. The AUC of 0.82 is indicative of an excellent/good classification of false positive and true positive. The FPR in this case was 0.01.**

### **Naïve Bayes With PCA**

**The same auto tuned parameters were chosen for the final model which yielded a Kappa of 0.01, an AUC of 0.59 and a FPR of 0.07. This is a model that doesn’t perform well when PCA is added.**

### **Regularization Models**

**The regularization models were not used with PCA because the coefficients would not be properly interpreted. In PCA, the coefficients represent the weights for the principal components instead of the original predictors. This would therefore be problematic. All three models were provided a grid from which the parameters could be auto tuned from.**

### **Lasso Model**

**With alpha set to 1 and lambda at 0.01, the Kappa was 0.30, the AUC at 0.89 and FPR at 0. It is important to point out that Lasso has the ability to inherently select features that it deems to have a higher predictability. As such, some features’ coefficients were shrunk to near zero. This would have been equivalent to manually removing some features that did not have a statistical difference with the outcome variable. Although the Kappa statistics indicates a fair agreement and is the highest so far, I suspect the model might be biased towards the majority classifier because of the zero FPR.**

### **Ridge Model**

**With alpha set to 0 and lambda at 0.4, the Kappa statistics was the same as Lasso at 0.30. This is a model that makes some coefficients very small without getting rid of them. Instead, the model will minimize the impact of the small coefficients if their associated features are irrelevant. At the core, it looks like the model fits the purpose of what I am trying to accomplish by deciding to keep all features. The AUC of 0.9 can be interpreted as the model having a 90% chance to distinguish between true and false positive. The FPR came up to 0.01.**

### **Elastic Net Model**

**A grid of parameters was provided for the auto tuned selection of alpha and lambda. A look at the auto tuned hyper parameter shows that the model used a combination of ridge and lasso because alpha was set to 0.5. Lambda in this case was 0.11. The Kappa statistics decreased to 0.23. The AUC was 0.86 while the FPR increased to 0.02.**

### **Ensemble Models**

### **Random Forest**

**The model was given a grid of parameters to chose from with mtry set to 2 for the final model. The Kappa statistics for the model turned out to be 0.54, the highest so far, indicating a moderate agreement between predicted and actual values. The AUC for the model is at 0.90 and the FPR is at 0.02. These numbers tell the picture of a model that was able to build many decisions trees based on a random subset of features. The final tree that is produced is the mode of the class**[[6]](#footnote-6)**.**

### **Gradient Boost**

**The auto tuned model returned the following hyper parameters for the final model: n.trees of 100, interaction.depth of 2, shrinkage of 0.1 and n.minobsinnode of 10. The resulting Kappa statistics was 0.27 which is one of the lowest. The AUC and FPR were respectively at 0.82 and 0.03.**

### **Ada Boost**

**The auto tuned parameters for the final model were nIter at 100 with method Real adaboost. This model did better than Gradient Boost with a Kappa of 0.39. This is the second highest Kappa so far. The AUC though was low at 0.69. The FPR was 0.02.**

### **Tree Bag**

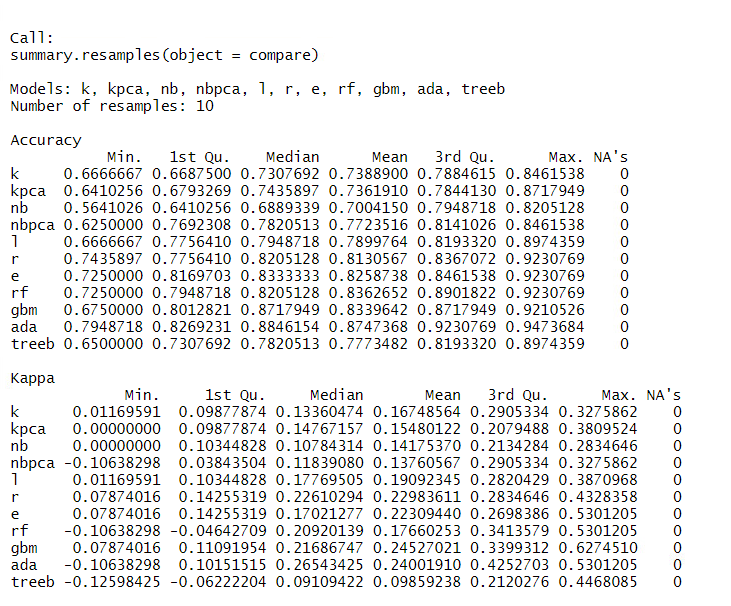
**As the simplest form of ensembling, the tree bag model did not perform well with a Kappa statistics of 0.21. Although its AUC is high at 0.84 and the FPR is 0.02.**

### **Neural Network**

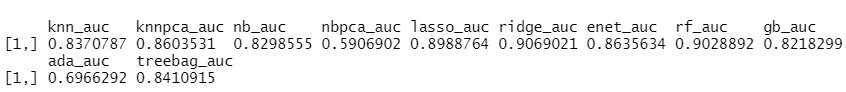
**A simple feedforward neural network was built with auto tuned parameters. The hyper-parameters of the model were tuned using a grid combination embedded in flags. Neural Networks are known to be complex models that work well with high dimensional data sets. It goes then that the model did not perform well on this low dimensional and highly imbalanced set because of the Kappa statistics of 0.09.**

## **Models Comparison**

**One of the first elements to notice is that all models did very well in terms of false positive rates. Using the resample function, I was able to compare the mean Kappa statistics of all models used so far. The model with the highest Kappa statistics, random forest, turns out not to have the highest mean for the in-sample data. AdaBoost has the second highest Kappa and the second highest mean. Ridge has the third highest Kappa and the third highest mean. The gap between the in-sample and out-of-sample Kappa is concerning and could mean that the random forest model is highly biased towards the majority classifier.**



**When it comes to the AUC metrics, Random Forest follows Ridge with the second highest values. If this were the only metrics, Random Forest would be without a doubt the chosen model for this data set.**



**Given the imbalanced nature of the outcome feature, the issue of bias within these metrics has to be seriously considered. I believe a model that performed in a balanced fashion for both the in-sample and out of sample data would be more appropriate. That model turns out to be the adaBoost model.**

## **Conclusion**

**The goal of this study was to build a model to predict whether a country is in recession or not. For such a task, it can be tempting to automatically choose the model with the best accuracy and lowest loss among others. I argue that this is not always the case especially when dealing with class imbalance. The results of the resample function for the highest performing model showed a discrepancy between the mean Kappa statistics for in-sample data and the Kappa statistics for out of sample data. This discrepancy could be due to the class imbalance. In that regard, a model that takes into account the information provided by weak learners might perform well. Another factor that should play into the selection of a model is the cost associated with the false positive and false negative rate. In this case, it is more desirable to have a yes recession case misclassified as no. In other words, the cost of a high false negative rate is better. Based on the observations from the Kappa statistics, Area Under Curve and taking into the account the issue of class imbalance, the AdaBoost model is what I would choose. I believe the model has a balanced performance across all metrics that still satisfies the ultimate goal of this project.**

# **References**

*African Country Recession Dataset (2000 to 2017).* (n.d.). Retrieved from Kaggle: https://www.kaggle.com/chirin/african-country-recession-dataset-2000-to-2017

Lantz, B. (2019). Random Forests. In B. Lantz, *Machine Learning with R: Expert techniques for predictive modeling* (p. 387). Packt Publishing Ltd.

Lantz, B. (2019). The Kappa Statistics. In B. Lantz, *Machine Learning with R: Expert techniques for predictive modeling* (p. 345). Packt Publishing Ltd.

*Recession detection in African countries.* (n.d.). Retrieved from Kaggle: https://www.kaggle.com/maistrishid/recession-detection-in-african-countries

Rhodes, D., & Stelter, D. (2009, February). *Seize Advantage in a Downturn.* Retrieved from Harvard Business Review: https://hbr.org/2009/02/seize-advantage-in-a-downturn

Rodeck, D., & Curry, B. (2021, March 28). *What is a Recession.* Retrieved from Forbes: https://www.forbes.com/advisor/investing/what-is-a-recession/

1. (Rodeck & Curry, 2021) [↑](#footnote-ref-1)
2. (Rhodes & Stelter, 2009) [↑](#footnote-ref-2)
3. (African Country Recession Dataset (2000 to 2017)) [↑](#footnote-ref-3)
4. (Recession detection in African countries) [↑](#footnote-ref-4)
5. (Lantz, The Kappa Statistics, 2019) [↑](#footnote-ref-5)
6. (Lantz, Random Forests, 2019) [↑](#footnote-ref-6)