SUPERVISED LEARNING: CLASSIFICATION

Khobie Maseko
14/12/2021

Project Background and Objective

- This data was obtained from a dataset by a Kaggle user named ‘chirin’ titled **African Country Recession Dataset (2000 to 2017**). The dataset 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. The blend is specifically created to answer the question: "**What factors contribute most to, or are most indicative of, recessions in Africa?**". It covers 27 African countries.
- Dateset source: https://www.kaggle.com/chirin/african-country-recession-dataset-2000-to-2017
- The dataset consists of 50 columns and 486 rows.
- For purposes of this project I will explore the relationship between the target (**growthbucket**) variable and potential predictors. I will train 3 classifier models.

Data Exploration Plan

The following preliminary steps are the methods we will utilise to attempt to build a baseline model to find out if there are any strong correlations between the different factors and the target variable:

1. Data Overview
2. Data Cleaning and Feature Engineering:
3. Classifier Model 1: Logistic Regression
4. Classifier Model 2: K-Nearest Neighbors
5. Classifier Model 3: Random Forest + Extra Trees
6. Insights and recommendations
7. Conclusion

Data Overview - Data Attributes

	Variable	Description
0	pop	Population (in millions)
1	emp	Number of persons engaged (in millions)
2	emp_to_pop_ratio	Ratio of Employed Persons to Total Population
3	hc	Human capital index, based on years of schooli...
4	ccon	Real consumption of households and government,...
5	cda	Real domestic absorption, (real consumption pl...
6	cn	Capital stock at current PPPs (in mil. 2011US$)
7	ck	Capital services levels at current PPPs (USA=1)
8	ctfp	TFP level at current PPPs (USA=1)
9	cwtfp	Welfare-relevant TFP levels at current PPPs (U...
10	rconna	Real consumption at constant 2011 national pri...
11	rdana	Real domestic absorption at constant 2011 nati...
12	rnna	Capital stock at constant 2011 national prices...
13	rkna	Capital services at constant 2011 national pri...
14	rtfpna	TFP at constant national prices (2011=1)
15	rwtfpna	Welfare-relevant TFP at constant national pric...
16	labsh	Share of labour compensation in GDP at current...
17	irr	Real internal rate of return
18	delta	Average depreciation rate of the capital stock
19	xr	Exchange rate, national currency/USD (market+e...
20	pl_con	Price level of CCON (PPP/XR), price level of U...
21	pl_da	Price level of CDA (PPP/XR), price level of US...
22	pl_gdpo	Price level of CGDPo (PPP/XR), price level of...
23	csh_c	Share of household consumption at current PPPs
24	csh_i	Share of gross capital formation at current PPPs
25	csh_g	Share of government consumption at current PPPs
26	csh_x	Share of merchandise exports at current PPPs
27	csh_m	Share of merchandise imports at current PPPs
28	csh_r	Share of residual trade and GDP statistical di...
29	pl_c	Price level of household consumption, price l...
30	pl_i	Price level of capital formation, price level...
31	pl_g	Price level of government consumption, price ...
32	pl_x	Price level of exports, price level of USA GDP...
33	pl_m	Price level of imports, price level of USA GDP...
34	pl_n	Price level of the capital stock, price level ...
35	total	Annual Bank of Canada commodity price index - ...
36	excl_energy	Annual Bank of Canada commodity price index - ...
37	energy	Annual Bank of Canada commodity price index - ...
38	metals_minerals	Annual Bank of Canada commodity price index - ...
39	forestry	Annual Bank of Canada commodity price index - ...
40	agriculture	Annual Bank of Canada commodity price index - ...
41	fish	Annual Bank of Canada commodity price index - ...
42	total_change	Year-on-Year Percentage Change Annual Bank of...
43	excl_energy_change	Year-on-Year Percentage Change Annual Bank of...
44	energy_change	Year-on-Year Percentage Change Annual Bank of...

Data Overview - Data Attributes cont'd

45	metals_minerals_change	Year-on-Year Percentage Change Annual Bank of...
46	forestry_change	Year-on-Year Percentage Change Annual Bank of...
47	agriculture_change	Year-on-Year Percentage Change Annual Bank o...
48	fish_change	Year-on-Year Percentage Change Annual Bank of...
49	growthbucket	"1" = Recession; "0" = No_Recession

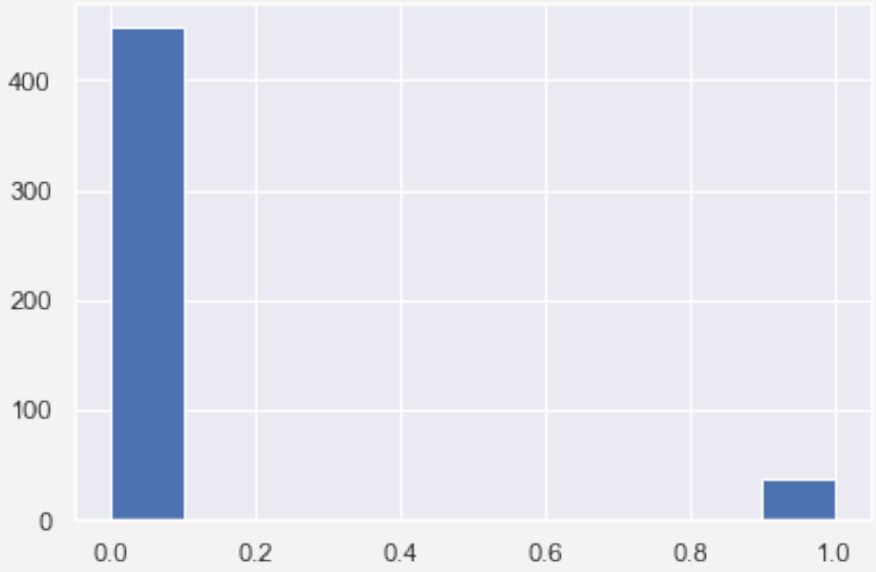
Data Overview - Data Description

	pop	emp	emp_to_pop_ratio	hc	ccon
count	486.000000	486.000000	486.000000	486.000000	486.000000
mean	20.185755	7.121089	0.357865	1.777389	64361.006942
std	30.037490	9.921471	0.080541	0.446339	129634.856793
min	1.061468	0.243000	0.198212	1.069451	2781.259277
25%	3.830730	1.048750	0.297922	1.445886	9117.209716
50%	10.868272	4.184000	0.368841	1.689902	17471.495120
75%	24.220695	8.517560	0.416717	2.117452	58016.873047
max	190.886307	65.156548	0.555433	2.885300	758455.187500

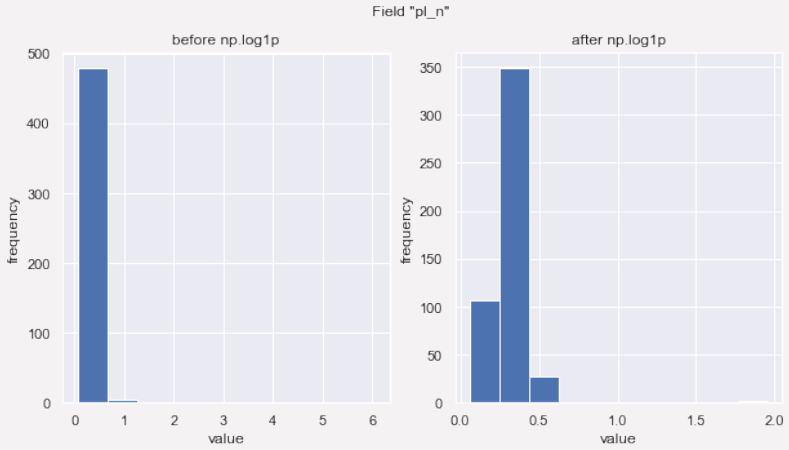
	cda	cn	ck	ctfp	cwtfp	...
count	486.000000	4.860000e+02	486.000000	486.000000	486.000000	...
mean	80885.988722	2.442244e+05	0.004583	0.454419	0.453431	...
std	156740.416624	4.725163e+05	0.008210	0.206562	0.203056	...
min	2984.366943	5.790397e+03	0.000124	0.098622	0.107790	...
25%	11081.697755	2.429231e+04	0.000514	0.301179	0.295615	...
50%	22228.022460	6.432356e+04	0.001355	0.400647	0.405870	...
75%	69676.791020	1.886244e+05	0.003227	0.616736	0.603459	...
max	896604.812500	2.886312e+06	0.041835	0.998187	1.031707	...

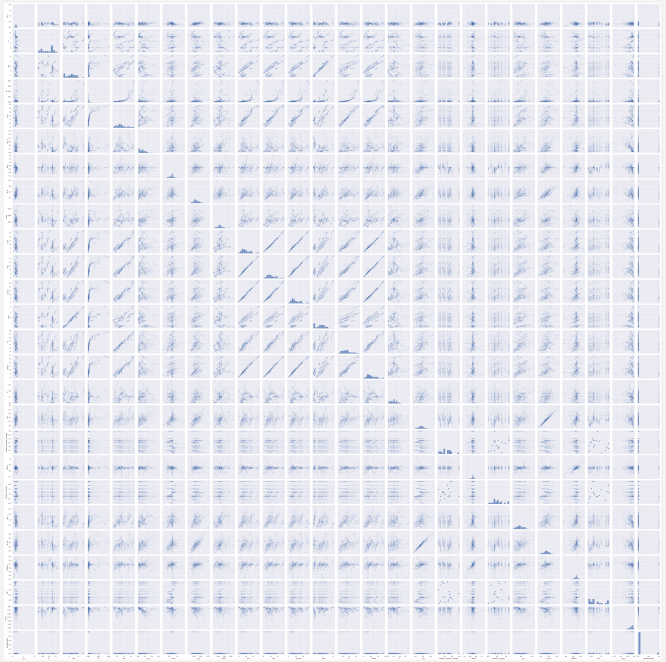
	agriculture	fish	total_change	excl_energy_change	\
count	486.000000	486.000000	486.000000	486.000000	
mean	221.326667	1009.445556	0.044535	0.030346	
std	49.803981	140.125951	0.184025	0.105616	
min	149.370000	843.660000	-0.359446	-0.173741	
...					
25%			-0.026451	-0.020995	0.000000
50%			0.038444	0.006529	0.000000
75%			0.112793	0.047421	0.000000
max			0.320880	0.165440	1.000000

To the left are the descriptive statistics of the dataset. We determined that our target variable is unevenly distributed using the below histogram.We noticed that all but one of the columns are float64. There are 50 columns and 486 rows in this dataset.



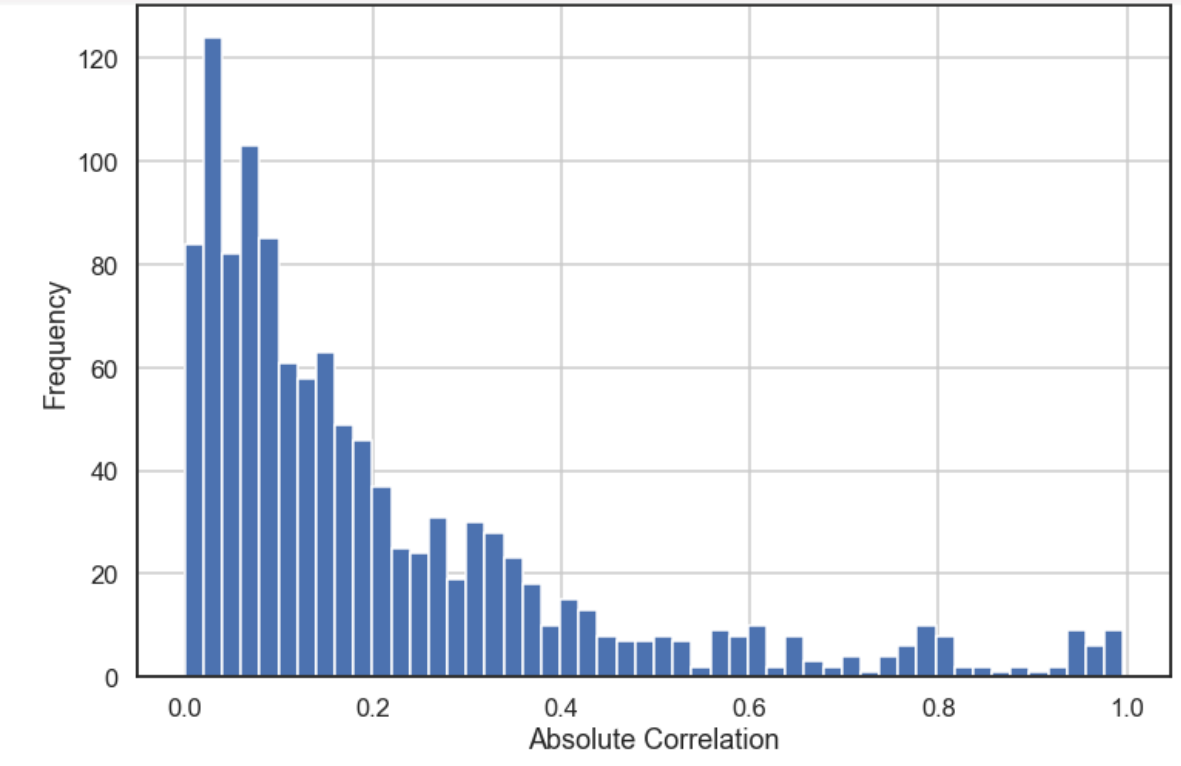
Data Cleaning and Feature Engineering





- We used Log Transform to adjust the skewed data, this decreased the effect of any outliers.
- We generated pairplot visuals to better understand the target and feature-target relationships.
- We saw that the target variable does not seem to have a linear relationship with any of the features.
- We applied basic feature transformations to the feature columns and saw a difference in them.

Data Cleaning and Feature Engineering



We then:

- calculated the correlation values;
- emptied all the data below the diagonal.
- Made the unused values NaN;
- nulled out all the values on the diagonal and below;
- recreated the correlation pandas dataframe;
- stacked the data and converted it to a data frame;
- made a histogram of the absolute value correlations.

- These are the most highly correlated values.
- Values are sorted by correlation going from top to bottom.
- We only wanted values from 0.8 onwards.
- We then split the data into Train/Test sets using **StratifiedShuffleSplit** to maintain the same ratio of predictor classes

Classifier Model 1: Logistic Regression

	lr	l1	l2
	0	0	0
15	-0.303797	0.0	-1.151993e-04
44	0.924955	0.0	1.684318e-04
9	-0.661641	0.0	-2.861378e-04
47	0.203679	0.0	9.972202e-06
17	0.039872	0.0	-9.416069e-06
18	0.019871	0.0	-8.873865e-07
31	-0.880191	0.0	-1.793979e-04
34	-0.521691	0.0	-1.299903e-04
5	0.086030	0.0	-9.298099e-04
32	-0.155033	0.0	9.697661e-06

- We fit a logistic regression model without any regularization using all of the features.
- We then made a regularized logistic regression models, L1 and L2.
- We used LogisticRegressionCV because it works like cv grid search.
- We combined all the coefficients into a dataframe.

Classifier Model 1: Logistic Regression

	lr	l1	l2
precision	0.924658	0.924658	0.924658
recall	0.924658	0.924658	0.924658
fscore	0.924658	0.924658	0.924658
accuracy	0.924658	0.924658	0.924658

We then predicted the class and the probability for each model.

For each model, we calculate the following error metrics: Accuracy, Precision, Recal, F-score and Confusion Matrix.

We noted that the error metrics are the same due to how we configured the Logistic Regression algorithm.

The percentage rate for this model is **92.46%** though, which is relatively high.

Classifier Model 2: K-Nearest Neighbors

	count	mean	std	min	25%	50%	75%	max
lr	49.0	0.009720	0.410856	-1.188556	-0.135816	0.021466	0.203679	0.924955
l1	49.0	-0.000039	0.000273	-0.001909	0.000000	0.000000	0.000000	0.000000
l2	49.0	-0.000239	0.000640	-0.003217	-0.000286	-0.000011	0.000025	0.001123

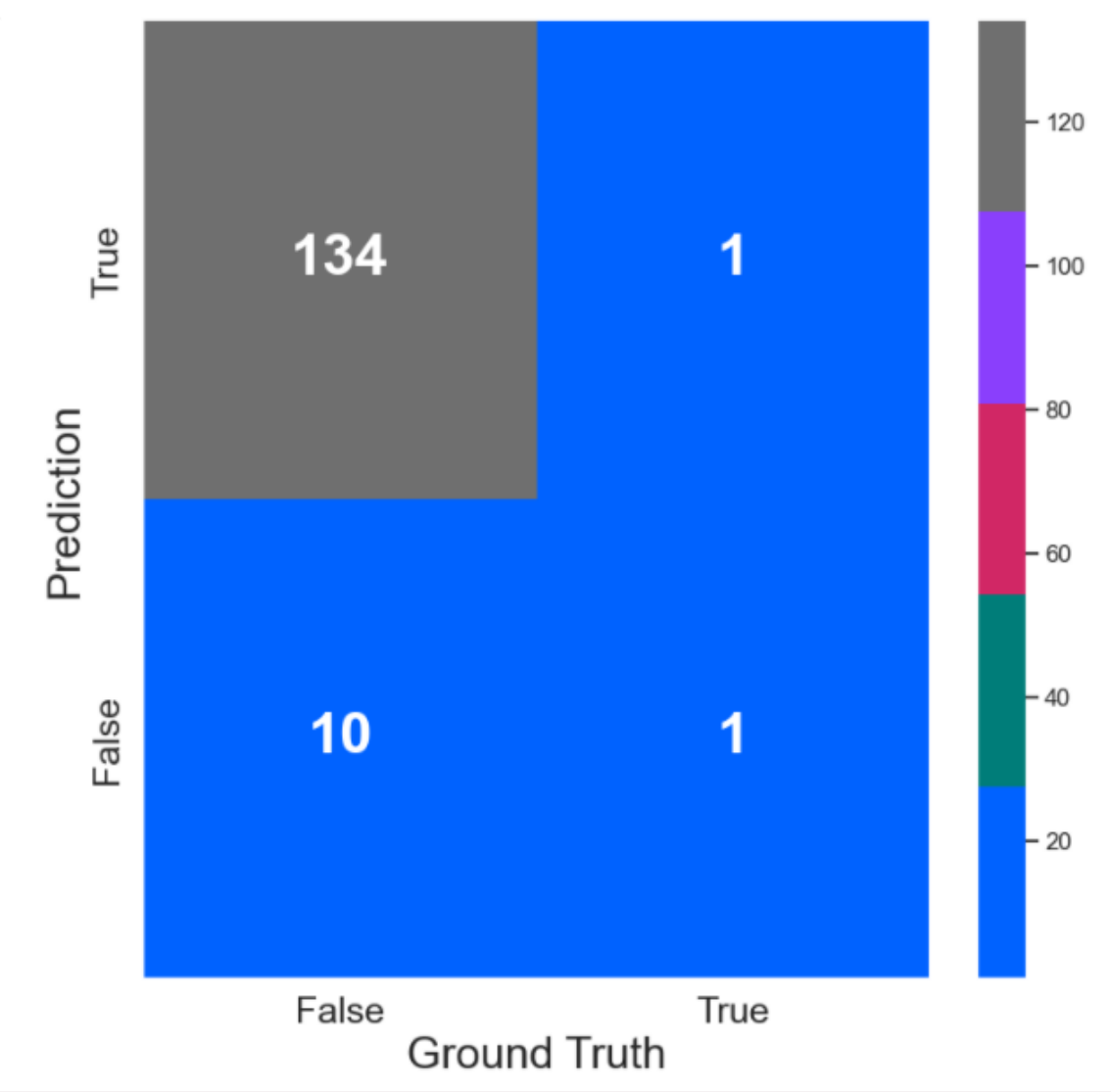
	count	mean	std	min	25%	50%	75%	max
lr	49.0	0.01	0.411	-1.189	-0.136	0.021	0.204	0.925
l1	49.0	-0.00	0.000	-0.002	0.000	0.000	0.000	0.000
l2	49.0	-0.00	0.001	-0.003	-0.000	-0.000	0.000	0.001

```
              precision    recall  f1-score   support

           0       0.93      0.99      0.96       135
           1       0.50      0.09      0.15        11

    accuracy                           0.92       146
   macro avg       0.72      0.54      0.56       146
weighted avg       0.90      0.92      0.90       146

Accuracy score:  0.92
F1 Score:  0.15
```

We noticed that the variables were on different scales so we scaled the data using the MinMaxScaler.

We also estimated the KNN model and report outcomes before calculating the Accuracy, Precision, recall, f-scores from the multi-class support function.

Classifier Model 2: K-Nearest Neighbors

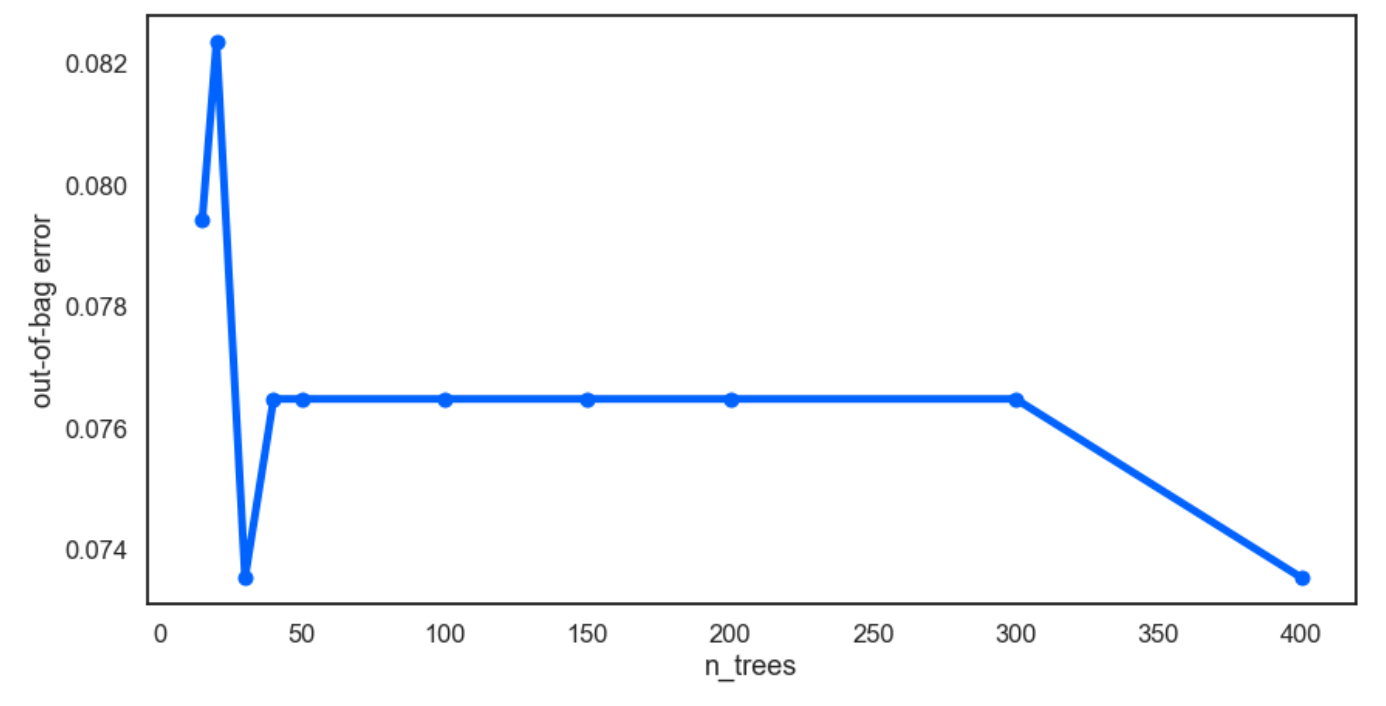


We noted that the accuracy score for this KNN model is **92.00%**, which is lower than the Logistic Regression model's score. The F1 score is a very low (15%).

The confusion matrix shows us that this model made 134 true predictions (out of a possible 135) and 10 false ones (out of a possible 11) on the test data.

The model got high predictions for the “0” class and really low predictions for the “1” class of the data. This is probably caused by the huge class imbalance that we previously noted.

Classifier Model 3: Random Forest



We iterated through all of the possibilities for n number of trees:

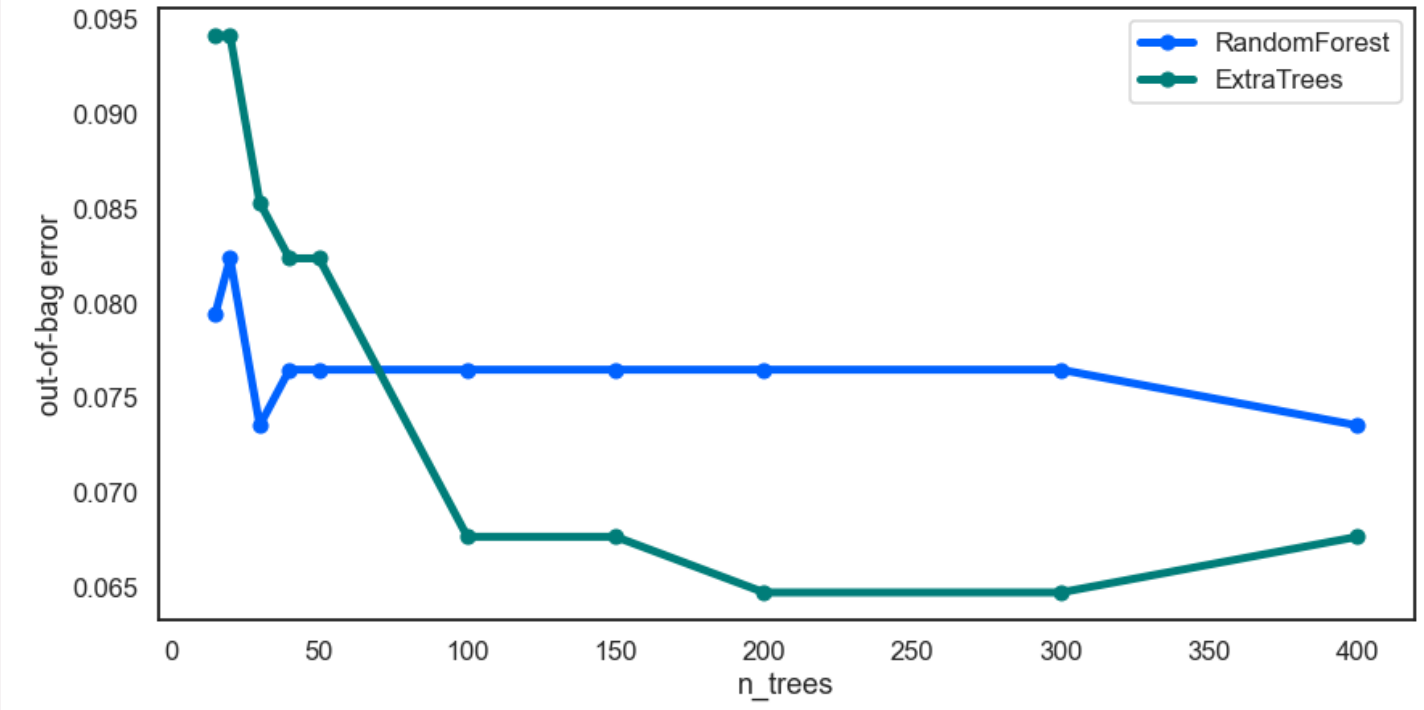
50,100,150,200,250,300,350 and 400.

We then:

- fit the model and
- got the out-of-bag error.

The error stabilized at around 40 trees. This is where it plateaus.

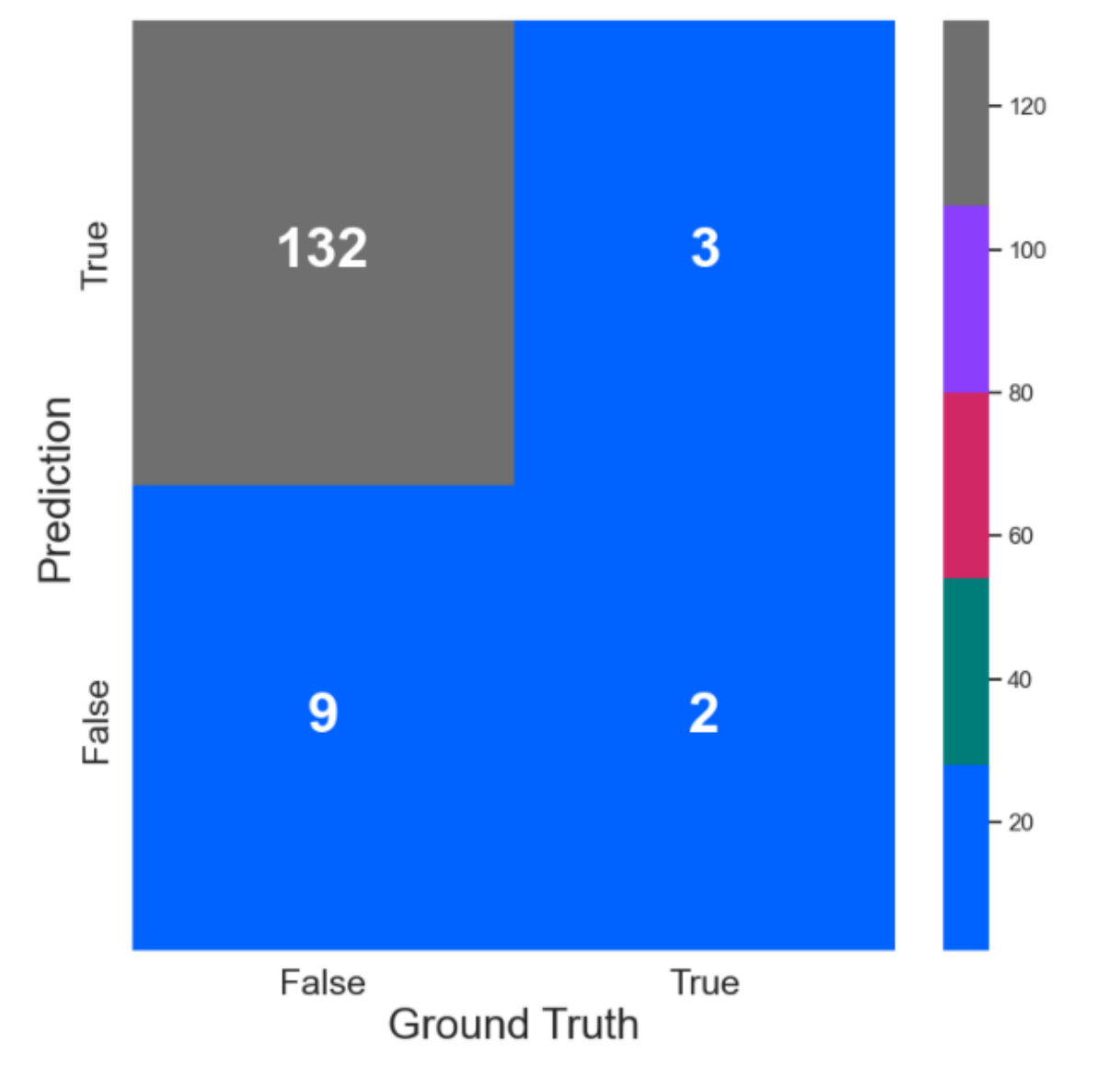
Classifier Model 3: Extra Trees



We initialized a extra random forest estimator (extra trees) for comparison and combined the two dataframes.

The error for this estimator stabilized at around 100 trees so it performed better than the Random Forest one.

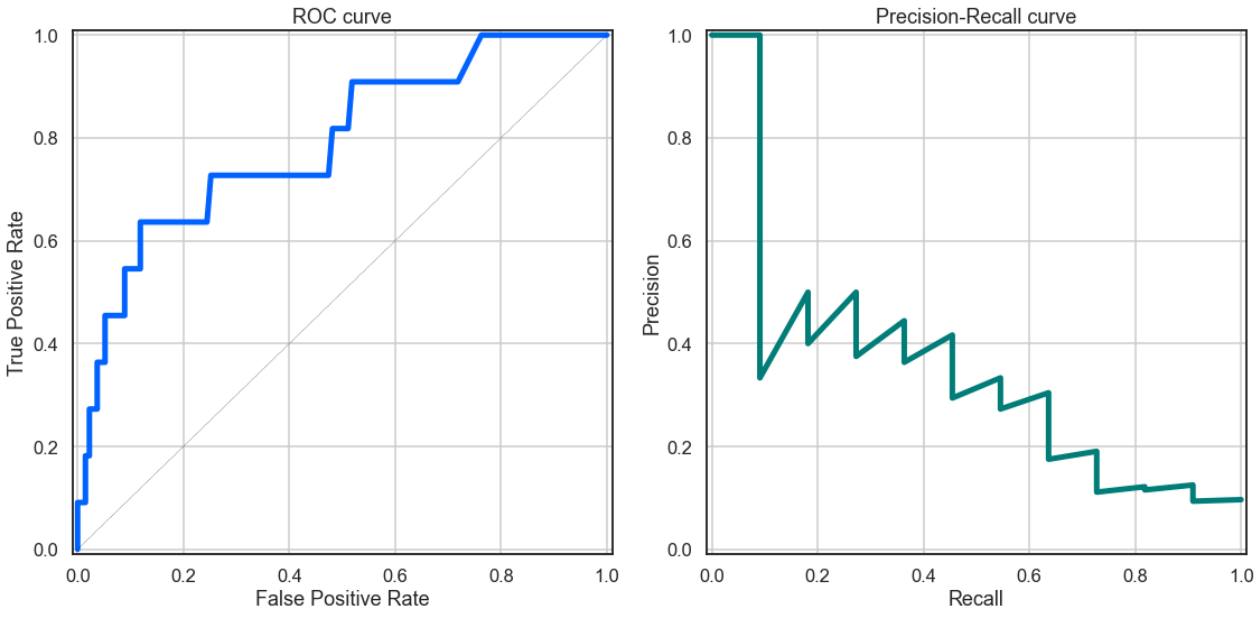
Classifier Model 3: Random Forest + Extra Trees



We noted that the precision, f1, support score and recall for "0" are high but low for "1" again.

The general accuracy score is **91.77%** which is lower than the Logistic Regression and the KNN models' scores.

Classifier Model 3: Random Forest + Extra Trees



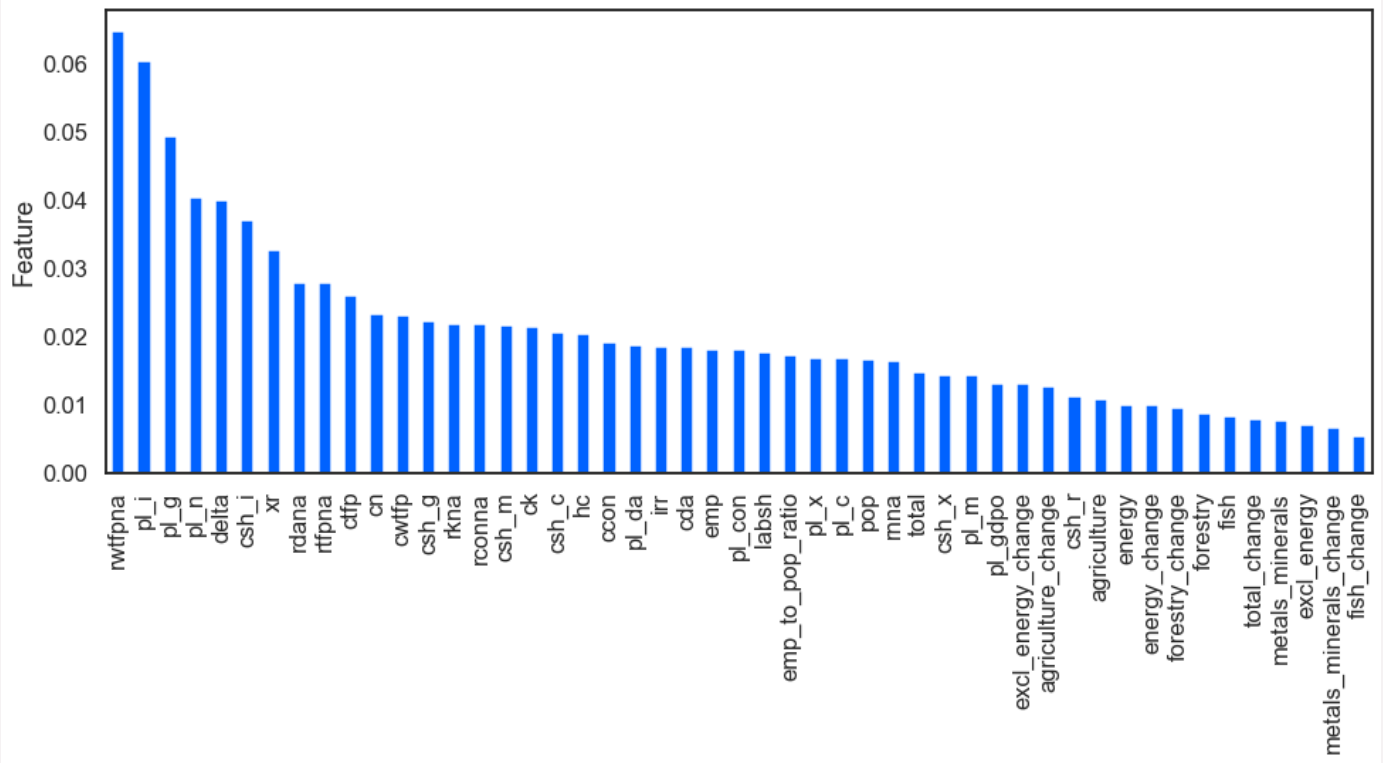
We also produced a ROC and a Precision-Recall curve to further explore the data.

We noted that the ROC curve is far from the top left corner so the test is not efficient. This makes sense since this data is imbalanced.

The Precision recall curve is also not efficient for this data. This is all due to the data's accuracy being low for the "1" class as noted before.

Insights

The bar chart below shows us which features are the most important when it comes to determining which features have the most influence on the target variable. In this case it is the "rwtfpna" (TFP at constant national prices (2011=1) feature. This feature is basically the portion of output that is not explained by the amount of inputs used in production for the reference year 2011.



Recommendations

I explicitly recommend that this dataset be feature engineered using a hybrid upsampling and downsampling method such as **SMOTE+TOMEK** for better prediction accuracy. The imbalance of the dataset does not allow for proper modelling. The use of a method such as S+T would also prevent overfitting of any models that are tested on this dataset.

I would also recommend obtaining same kind of data from the other 27 African countries that have been excluded in this dataset for a more thorough insight into the objective of this project.

Conclusion

With all that being said, we can conclude that the best model to use on this dataset (prior to the use of a method such as SMOTE + TOMEK) is the Logistic Regression model. It has the highest accuracy score, although not by much as compared to the KNN and Random Trees models.

Jupyter Notebook

The Jupyter Notebook for this project can be viewed at:

https://github.com/KhobieMaseko/IBM-Machine-Learning-Professional-Certificate-/blob/1a04443746da4b2627c722fe35e9a7db72e866aa/Project%203%20Supervised%20Learning%20-%20Classification1.ipynb