SUPERVISED LEARNING: REGRESSION

Khobie Maseko
16/11/2021

Project Background

- 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 clean the data and explore the relationship between the target (**growthbucket**) and potential predictors.

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. Cross Validation
4. Polynomial Features
5. Regularization
6. Insights and recommendations
7. Conclusion

Data Overview (1) - 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 (2) - 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 (3) - Data Columns

```
['pop', 'emp', 'emp_to_pop_ratio', 'hc', 'ccon', 'cda', 'cn', 'ck', 'ctfp', 'cwtfp', 'rconna', 'rdana', 'rnna', 'rkna', 'rtfpna', 'rwtfpna', 'labsh',
'irr', 'delta', 'xr', 'pl_con', 'pl_da', 'pl_gdpo', 'csh_c', 'csh_i', 'csh_g', 'csh_x', 'csh_m', 'csh_r', 'pl_c', 'pl_i', 'pl_g', 'pl_x', 'pl_m', 'pl_n',
'total', 'excl_energy', 'energy', 'metals_minerals', 'forestry', 'agriculture', 'fish', 'total_change', 'excl_energy_change', 'energy_change',
'metals_minerals_change', 'forestry_change', 'agriculture_change', 'fish_change', 'growthbucket']
```

- All the different columns in the dataset
- Our target variable is '**growthbucket**'

Data Overview (4) - Data Types

```
pop                          float64
emp                          float64
emp_to_pop_ratio             float64
hc                           float64
ccon                         float64
cda                          float64
cn                           float64
ck                           float64
ctfp                         float64
cwtfp                        float64
rconna                       float64
rdana                        float64
rnna                         float64
rkna                         float64
rtfpna                       float64
rwtfpna                      float64
labsh                        float64
irr                          float64
delta                        float64
xr                           float64
pl_con                       float64
pl_da                        float64
pl_gdpo                      float64
csh_c                        float64
csh_i                        float64
csh_g                        float64
csh_x                        float64
csh_m                        float64
csh_r                        float64
pl_c                         float64
pl_i                         float64
pl_g                         float64
pl_x                         float64
pl_m                         float64
pl_n                         float64
total                        float64
excl_energy                  float64
energy                       float64
metals_minerals              float64
forestry                     float64
agriculture                  float64
fish                         float64
total_change                 float64
excl_energy_change           float64
energy_change                float64
metals_minerals_change       float64
forestry_change              float64
agriculture_change           float64
fish_change                  float64
growthbucket                   int64
dtype: object
```

- All the different data types of the dataset.
- We notice that all but one of the columns are float64.
- The target variable column is a integer datatype.

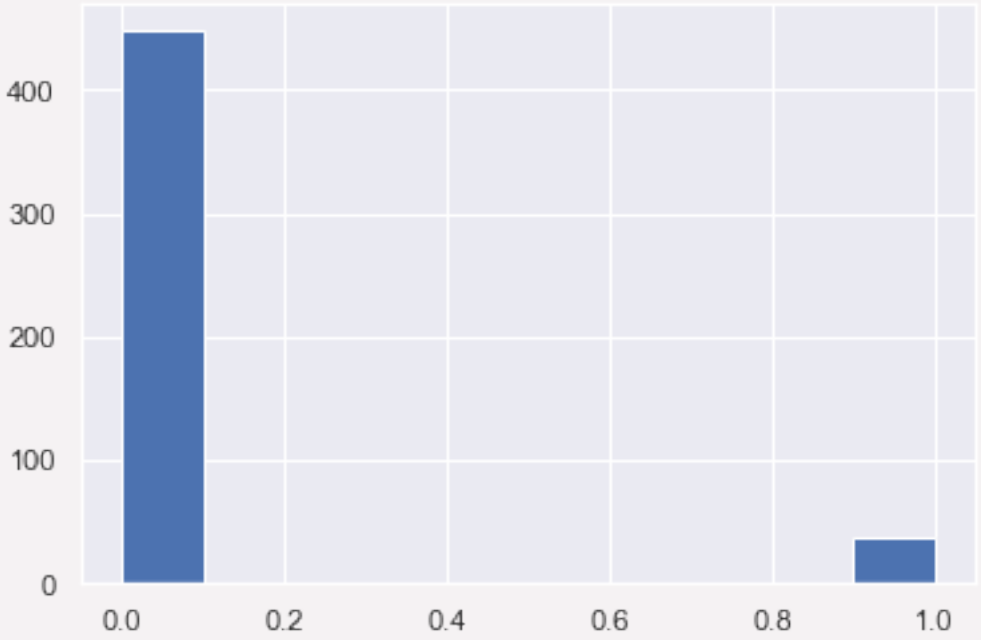
Data Overview (5) - Missing Values

```
pop                          0
emp                          0
emp_to_pop_ratio             0
hc                           0
ccon                         0
cda                          0
cn                           0
ck                           0
ctfp                         0
cwtfp                        0
rconna                       0
rdana                        0
rnna                         0
rkna                         0
rtfpna                       0
rwtfpna                      0
labsh                        0
irr                          0
delta                        0
xr                           0
pl_con                       0
pl_da                        0
pl_gdpo                      0
csh_c                        0
csh_i                        0
csh_g                        0
csh_x                        0
csh_m                        0
csh_r                        0
pl_c                         0
pl_i                         0
pl_g                         0
pl_x                         0
pl_m                         0
pl_n                         0
total                        0
```

```
excl_energy                  0
energy                       0
metals_minerals              0
forestry                     0
agriculture                  0
fish                         0
total_change                 0
excl_energy_change           0
energy_change                0
metals_minerals_change       0
forestry_change              0
agriculture_change           0
fish_change                  0
growthbucket                 0
dtype: int64
```

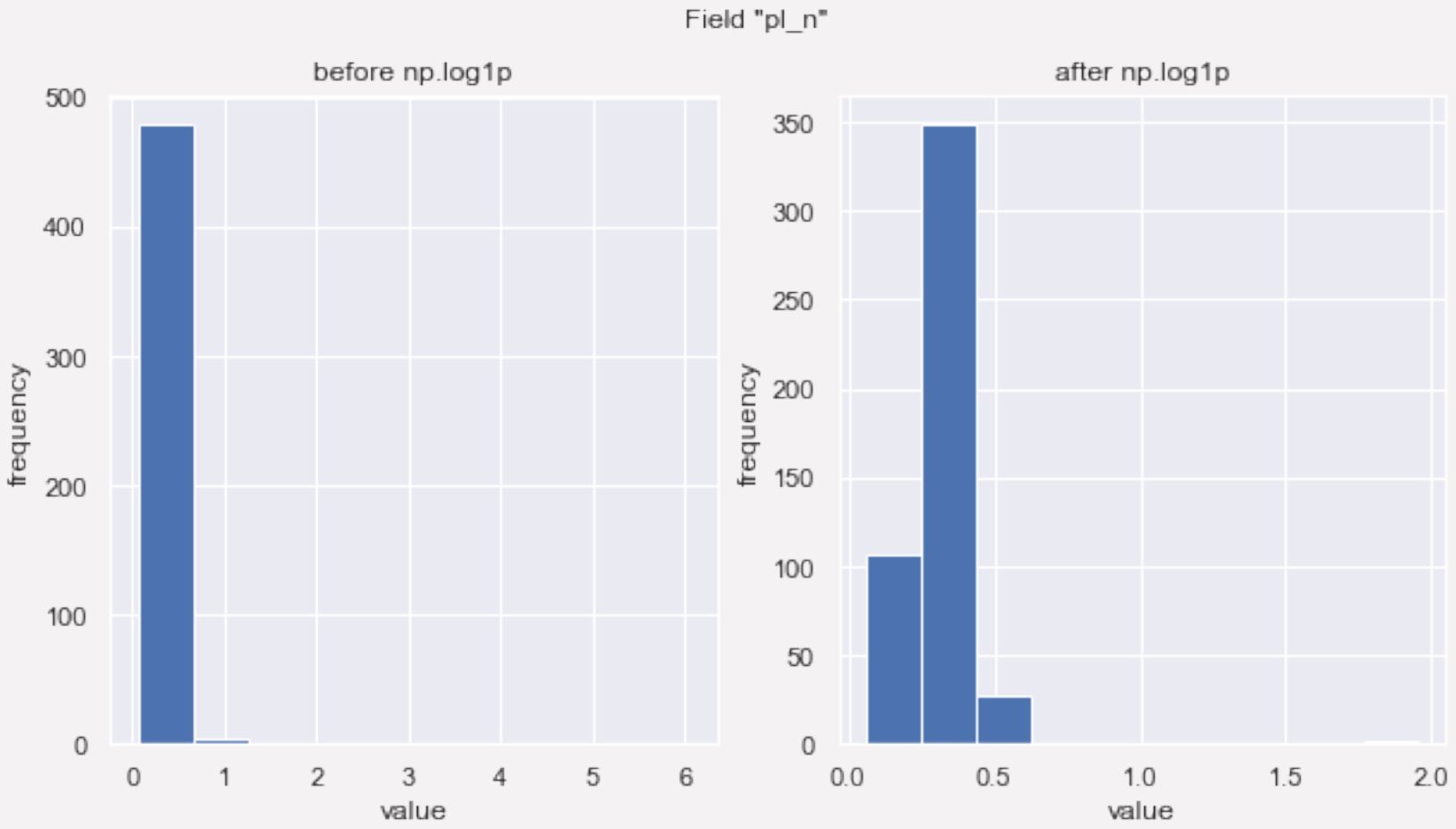
- We check for any missing values that might be in the dataset.
- We conclude that there are no missing values in the dataset.

Data Feature Engineering (1) - Histogram



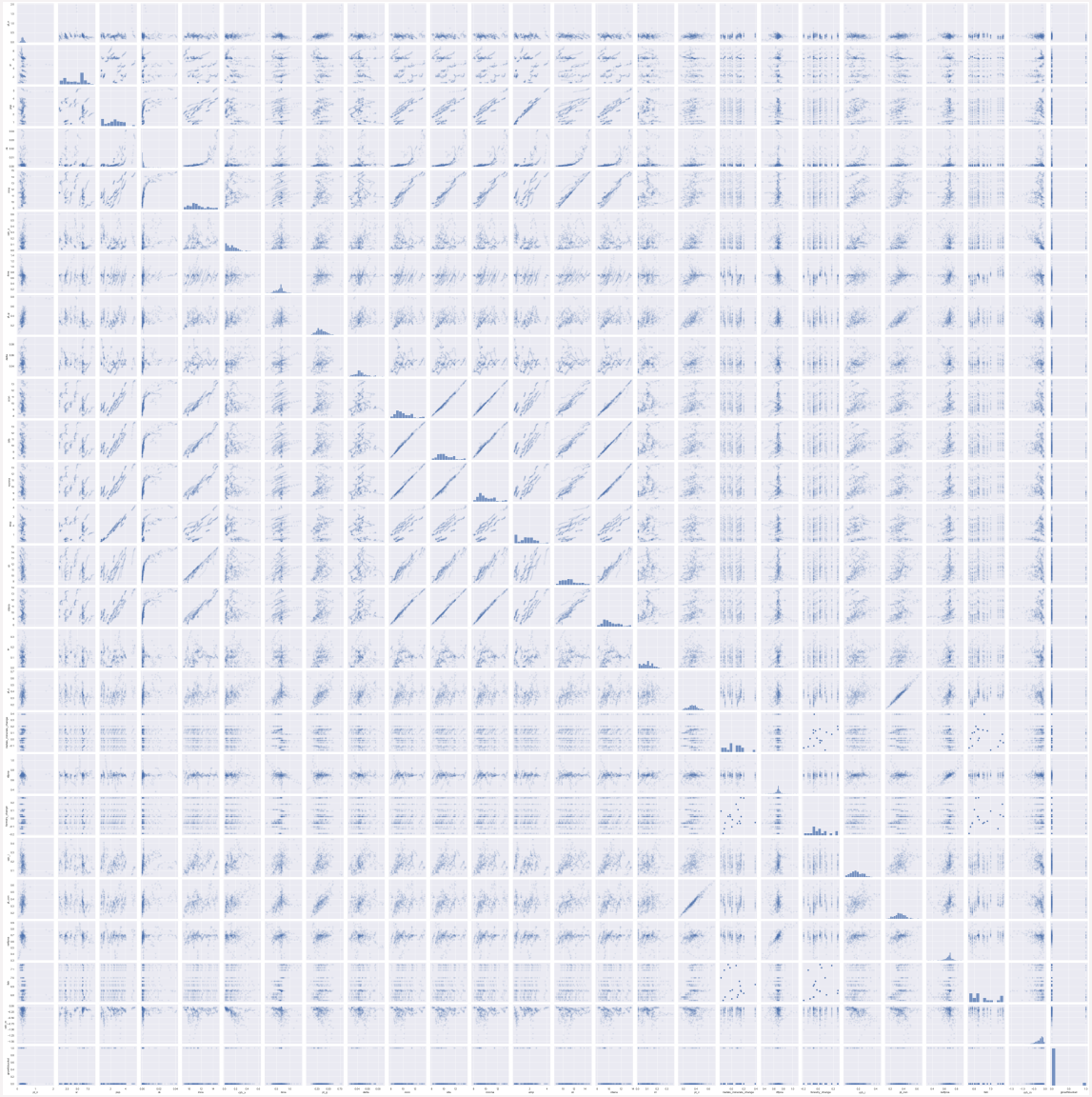
- We determined if our target variable is normally distributed using a histogram
- The histogram looks the way it does due to the categorical nature of our target variable, this is not surprising.
- We also note that most of the countries on the dataset have a "0" denotation, meaning that a majority of the countries were not considered to be in recessions during the documented period.

Data Feature Engineering (2) - Numerical Data



- We used Log Transform to handle any possible skewed data. It decreases the effect of the outliers, due to the normalization of magnitude differences and the model becomes more robust.
- We can see that the skew transformation on one of the features makes a difference to features that need it

Data Feature Engineering (3) - Numerical Data cont'd



- There appears to be no NaN values in the data set. Our dataset was perfectly filtered.
- We generated pairplot visuals to better understand the target and feature-target relationships.
- We can see that the target variable does not seem to have a linear relationship with any of the features.

Data Feature Engineering (4) - MSE

	Mean Squared Error
train	0.050026
test	0.070545

- We created train and test splits of the data sets.
- For each data set, fitted a basic linear regression model on the training data.
- We then calculated the mean squared error on both the train and test sets for the respective models.
- We noted that the MSE of train set is lower than that of the test set.
- We also noted that both MSEs are below 0.1 so they seem to be in a acceptable range at first glance.

Cross Validation - KFold

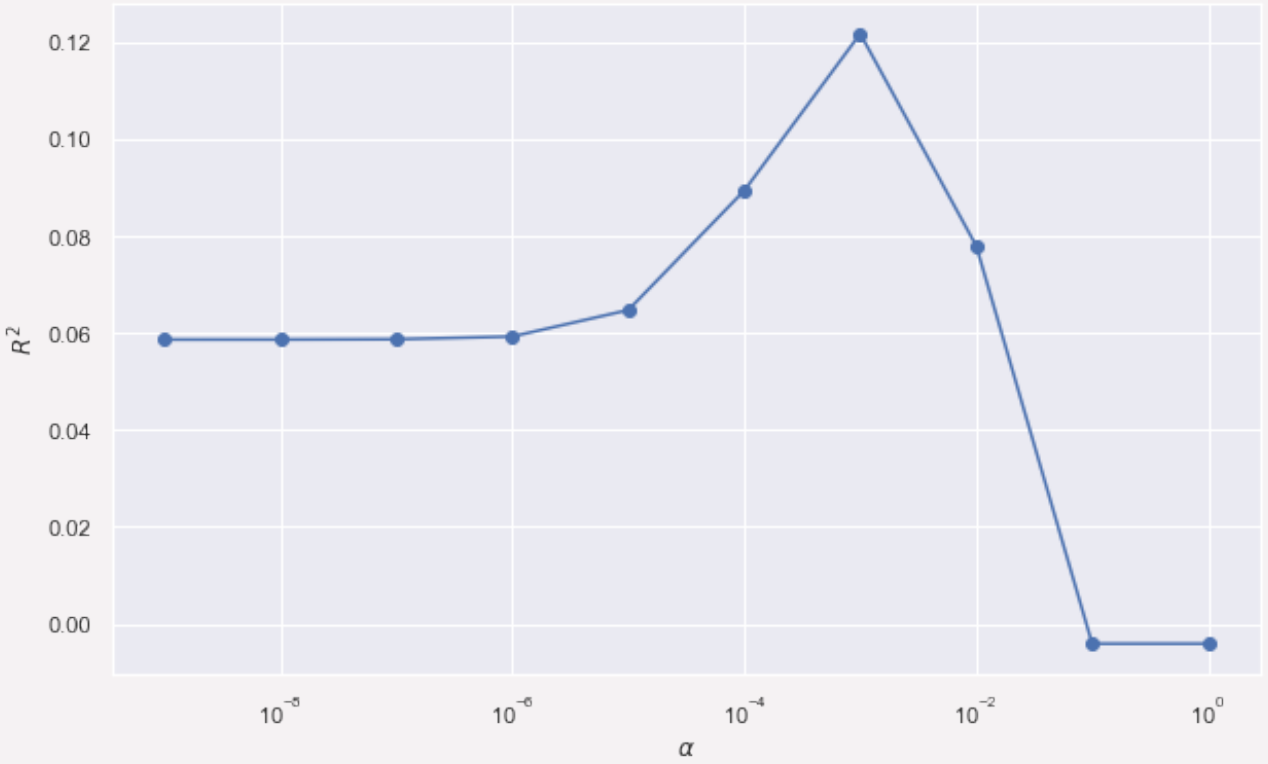
```
Train index: [ 1  3  4  5  7  8 10 11 12 13] 324
Test index: [ 0  2  6  9 15 17 19 23 25 26] 162

Train index: [ 0  2  6  9 10 11 12 13 15 17] 324
Test index: [ 1  3  4  5  7  8 14 16 22 30] 162

Train index: [0 1 2 3 4 5 6 7 8 9] 324
Test index: [10 11 12 13 18 20 21 24 28 31] 162
```

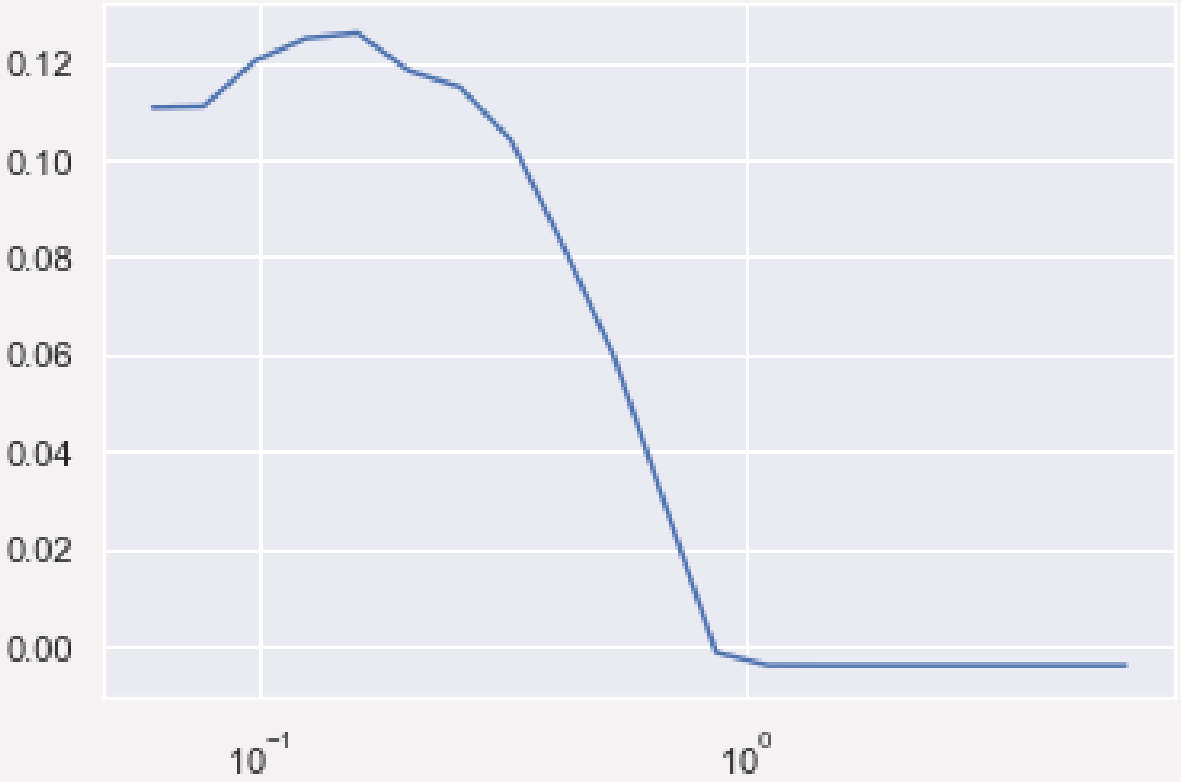
- To get a more thorough picture we used the `KFolds` object to split data into multiple folds (in this instance - 3 folds)
- Then we chained multiple data processing steps together using `Pipeline` to train three linear regression models: Linear, Lasso and Ridge.
- We used StandardScaler in the pipeline for cross validation.
- We then used cross_validation_predict to do K-fold cross validation for us.
- The r2 score was in a acceptable range: 0.058684.
- The estimator was still not fitted though.

Cross Validation - Hyperparameters



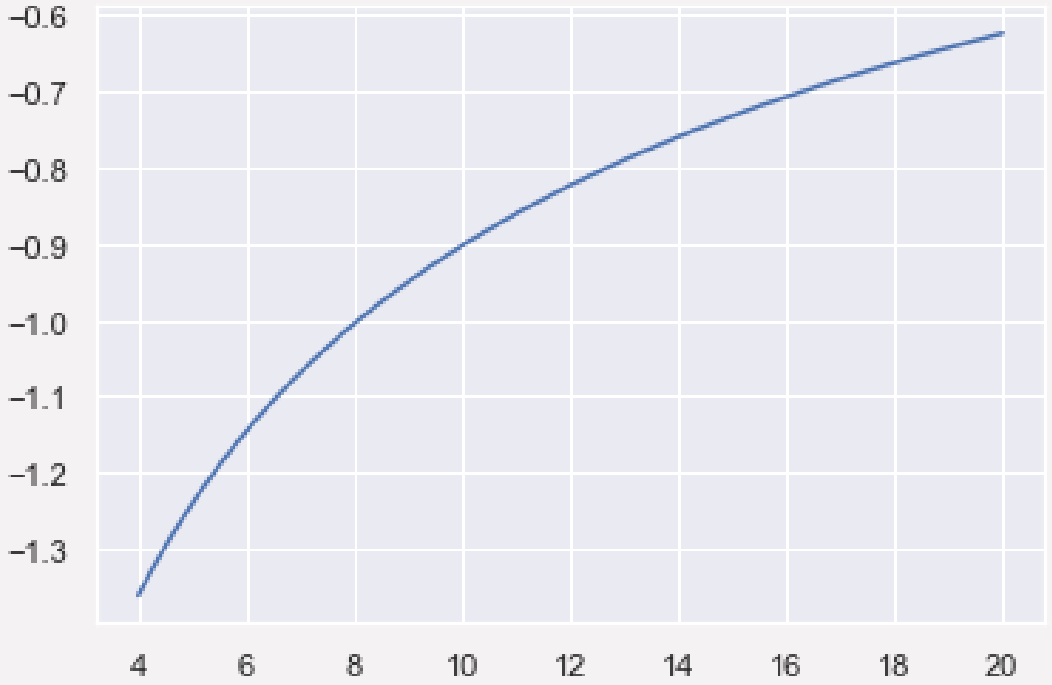
- We used Hyperparameter Tuning to determine which hyperparameters are most likely to generate a model that generalizes well outside of our sample.
- We tuned the `alpha` parameters we genrated for Lasso regression and plotted the results on the left.
- The results from the plot were underwhelming.

Polynomial Features(1) - Lasso Regression



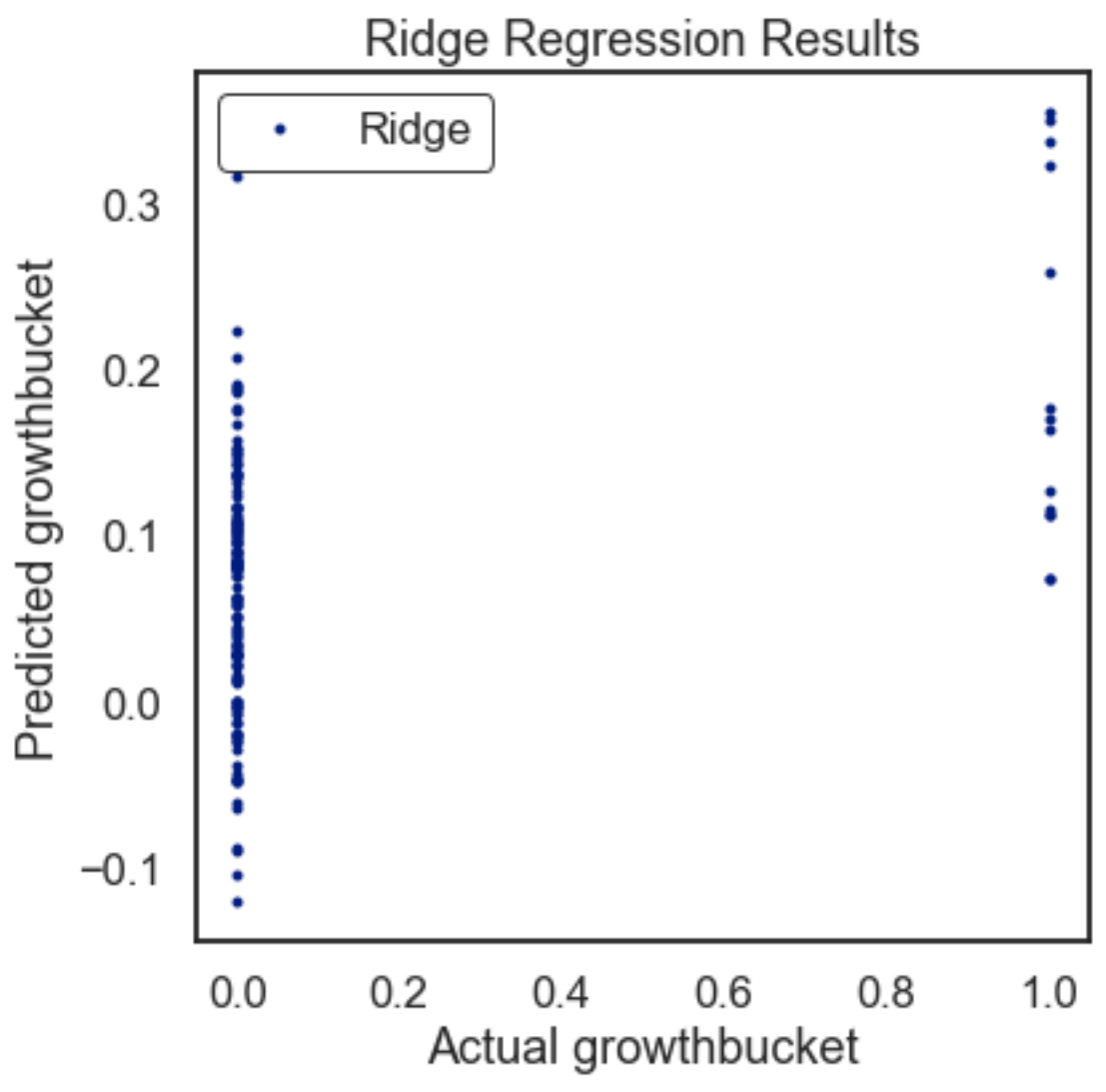
- We added Polynomial Features to the Pipeline and re-ran the cross validation.
- This cross validation is still for the Lasso regression model.
- We used the calculated parameters to try this model and obtained a estimation score of 0.44005181642058677 which is very low.

Polynomial Features(2) - Ridge Regression



- We tested the Ridge regression model in the same manner using Polynomial Features.
- This model gave us a estimator score of a whooping 0.99995857292867.
- We decided to use this model for this dataset.

Regularization



- We regularized the Ridge regression model due to its very high estimator score.
- We determined how many of the features remained non-zero: 49.
- The plot on the left shows us the predicted target variable vs the actual one.

Insights and recommendations

I would recommend that the Ridge regression model not be used for this dataset's target variabl. although the estimator score is 99% as opposed to the the 44% estimated from the Lasso model. I, however, made a note that the target variable on the dataset has a categorical nature so the use of a regression model to determine which feature(s) affect(s) it greatly is rendered moot.

Conclusion

This analysis has shown us that linear regression might not be a good fit for this particular dataset's target variable.It would be best tested using categorical models, more especially the Cost-Sensitive Classification.

Jupyter Notebook

The Jupyter Notebook for this project can be viewed at:

https://github.com/KhobieMaseko/IBM-Machine-Learning-Professional-Certificate-/blob/1e6e95ac2c7b98149bda21087dc6586e1be66b60/Project%202%20Supervised%20Learning%20-%20Regression%20%5BAfrican%20Country%20Recession%20(2000%20to%202017)%5D.ipynb