Capstone Project

Isac S. Camara Jr.

Machine Learning Engineer Nanodegree

September 21th, 2018

Definition

Project Overview

GDP per capita is a macroeconomic measurement used to compare and evaluate living in different countries, comparisons are made usually on the basis of nominal GDP but for comparatives in living standard PPP is used.

In this project, I created a predictive model for GDP per capita using two sources of data: a macroeconomic dataset found in Kaggle and the Heritage Economic Freedom Index. Time series are used in the Heritage Economic Freedom Index as a different data point to solve the lack of data.

The final model can predict GDP per capita using the Heritage Economic Freedom Index Raw data one can use the model to further understand how each features affect the GDP per capita and how government can improve living and economic wealth expending less attacking not the areas that have worse performance but the ones that can generate more GPD per capita improvement.

The motivation is find how the State affect individual lives, which public action can generate more financial health for the given country inhabitants, create a tool to show where to improve. In a more scientific language having “x” monetary unit that can generate a delta of “y” in a “z” feature find the argmax for GDP per capita. In a simple language where to put to money.

In the paper: New Tools for Predicting Economic Growth Using Machine Learning A Guide for Theory and Policy, one can read about machine learning being applied to this domain area, using decision trees and ensemble methods is this paper the researchers used up 39 and 20 features to predict the growth.

https://www.researchgate.net/publication/291827961_New_Tools_for_Predicti
ng_Economic_Growth_Using_Machine_Learning_A_Guide_for_Theory_and_Policy

Problem Statement

Having the given features predict GDP per capita, how well is GDP per capita explain by econometrics and how well is by a freedom index. How correlated is economic freedom to GDP per capita and how it is to the econometrics. Having x monetary unit that can generate a delta of y in a z feature find the argmax for GDP per capita.

The purpose is to construct one model that can be used to better understand GDP per capita and how economic freedom affects it.

1. Download the dataset of kaggle and the Heritage Economic Freedom Index
2. Clean the data and prepossessing for usage
3. Train a regressor to predict GDP per capita
4. Use the model to transform data into information

Metrics

Mean absolute percentage error and root mean squared error are regularly used to evaluate forecast methods. In this project I used the percentage of error to compare the models and the sources of data.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$$RMSPE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{A_t - F_t}{A_t}\right)^2}$$

Where A_t is the actual value and F_t is the forecast value.

This metrics was used when evaluating the models and when comparing with the benchmark model because the mean absolute error and mean squared error are not suitable for the problem tasks due the GDP per capita variance is high its vary more than a hundred times for the poorest country to the richest one. So a error of $1000 dollar are very high for the poorer ones and acceptable thought middle income and above countries.

And for the future users of the model the lack of precisions can:

- Led to bad assumptions in features importance
- Sub estimate one or other feature.
- Sub estimate freedom and liberty in the composition of GDP per capita.
- Led to bad decision in governments and public expenditure.

Analysis

Data Exploration

The kaggle dataset used can be found in: https://www.kaggle.com/sudalairajkumar/undata-country-profiles it has UN data over 229 countries and data about social, economic, environment. There are data regarding how production are configured and spaced and about educational measurement there are data over school enrollment. About health system in this data set of UN there number of physicians per 100 inhabitants and the government expenditure in it.

The kaggle dataset of UNdata used have this features:

1. Country – the country name.
2. Region – the region of the globe where the country is situated.
3. Surface area – the area of the country in km².
4. Population – the number of inhabitants that are living in the country.
5. Population density – population divided by surface area.
6. Sex Ratio - The difference in number of male and female inhabitants.
7. GDP – Gross Domestic Production in current US$ in millions.
8. GDP growth rate – the last year growth in GDP.
9. GDP per Capita – GDP divided by population our target variable.
10 11 & 12 Economy – Agriculture | Industry | Services (% GVA)

13, 14, 15. Employment - Agriculture | Industry | Services (% of employed)
16. Unemployment (% of labour force)
17. Labour force participation (female/male pop. %)
18, 19. Food and Agricultural Production Index
20,21,22: International Trade – Import | Export | Balance
23. Balance of Payment
24. Population Growth Rate
25. Urban Population % of total
26. Urban Population Growth Rate
27. Fertility Rate
28. Life Expectancy at birth
29. Population Age Distribution (0-14/ 60+)
30. Internacional Migrant Stock
31. Refugees and others of concern to UNHCR (in thousands)
32. Infant mortality rate (per 1000 live births
33, 34. Health: Expenditure | Physicians per 100
35,36,37,38. Education: Government expenditure |Primary | Secondary | Tertiary Gross Enroll
39. Seats held by women in national parliaments %
40. Mobile-cellular subscriptions
41. Individuals using the Internet
42. Threatened species
43. Forested area (% of land area)
44. CO2 emission estimates (million tons/tons per capita)
45,46 Energy production, primary (Petajoules) | Supply
47,48. Pop. Using improved drinking water | sanitation services.
49. Net Official Development Assist. received (% of GNI)

Not all of them were used to made the benchmark model, some of the features had no actual data or even no data at all, some data was manually collected from World Bank web site and CIA World Fact Book among others sources, so the models will use 47 features of this data set that have 229 countries, most of the data is continuous but one are categorical: region. That one will be treated and transformed to binary data leaving the data set with 47 + number of regions features, that after analysis became 70.

The Heritage data set can be found in the Heritage Economic Freedom Index web site: https://www.heritage.org/ , only data from 2013 forward were used, there are 20 years of data but the data from 2012 and backward cannot be directed comparable because the methodology used by heritage.

The heritage provides yearly a dataset with raw data, into it 186 data point for 2018 with twelve central data, the economic freedom index, that is divided in four areas:

1. Rule of Law
    1. Property Rights, - How property is respected is the given country
    2. Government Integrity - How much corruption is present in the everyday living.
    3. Judicial Effectiveness – How much time and how effective is the judicial system of the given country
2. Government Size
    4. Government Spending - it use a formula to transform general spending in percentage of GDP in a number of the index
    5. Tax Burden - the amount of taxation that the populations must have to pay in general
    6. Fiscal Health – how balanced is in spending and collecting.
3. Regulatory Efficiency
    7. Business Freedom – it measure the regulation efficiency in this given area for the given country, how it make business easy or hard to do.
    8. Labor Freedom – it measure the regulation efficiency in work contracts how easy and fair is to contract and fire employees
    9. Monetary Freedom – its measure the price stability
4. Open Markets
    10. Trade Freedom - is a composite measure of the extent of tariff and nontariff barriers that affect imports and exports of goods and services
    11. Investment Freedom – measure the regulation from investment in general in the given country
    12. Financial Freedom – measure banking efficiency and independency from government control

And some macroeconomic:

13. Tariff Rate – the average of taxation over the country economy
14. Income Tax Rate – the average of taxation inhabitants pays only for their yearly incomes.
15. Corporate Tax Rate – the average of percentage of tax over companies.
16. Government Expenditure – the amount of expenditures that the government does in a fiscal year, including staff
17. Population – number of inhabitants
18. GDP (PPP) -
19. GDP Growth Rate – percentage of change in GDP
20. 5 Year GDP Growth Rate – mean of growth in 5 years
21. GDP per Capita – the average income per year that each inhabitant has access to.(GDP is more complex than income but it a measurement of it.
22. Unemployment – the percentage of the economic active population that are searching for jobs and not finding any.
23. Inflation – the percentage of average change in prices
24. FDI Inflow – investment coming from others countries in millions of US dollar
25. Public Debt – how much the government have in debt

For information about the features used, how it are actually calculated what measurement and how each one of the affect the target variable and them the economy of the given country can be found in the heritage site in the section about the index and in the book in the section methodology, this section of the book can be also accessed and read in:
https://www.heritage.org/index/book/methodology

Exploratory Visualization

The plots below show the correlation between features the first and the second(Fig. 1 and Fig. 2) are a heat map of the correlations of the features in the heritage dataset and kaggle one, the third(Fig.3) is a matrix of plot in the diagonal, the distribution of the freedom index features among countries, outside the diagonal, pairwise relationship in the economic freedom index.

In this graphs (Fig.3) one can see the expected high correlation between the rule of law features and GDP per Capita but some interesting information tariff rate are high correlated with world rank and world rank with GDP, regions alone are not so correlated with GDP per capita as the region rank is. So have economic free neighbors is more important to be in the certain place, maybe because have economic free neighbors increases the commerce opportunities.

The scatters plot show the the are amost no nation with high labor and freedom from corruption that have low GDP per capita. In the second another interresting information employment in services area has a high correlation with GDP per capita than uneployment, and educational expenditure has amost no correlation with it.

Fig. 1

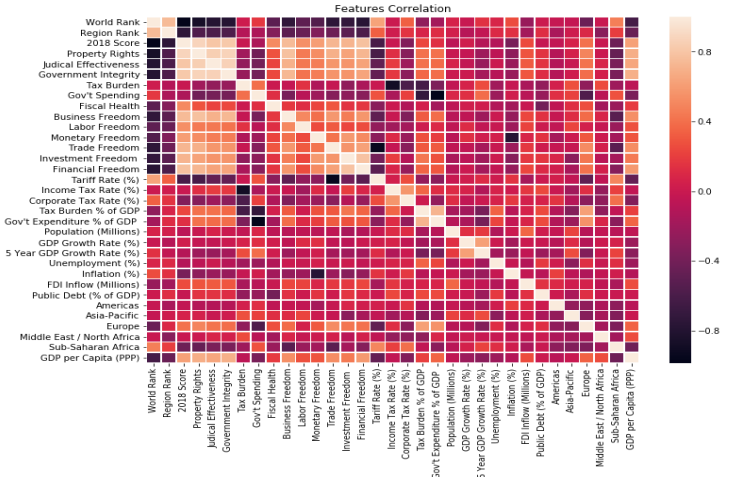


Fig. 2

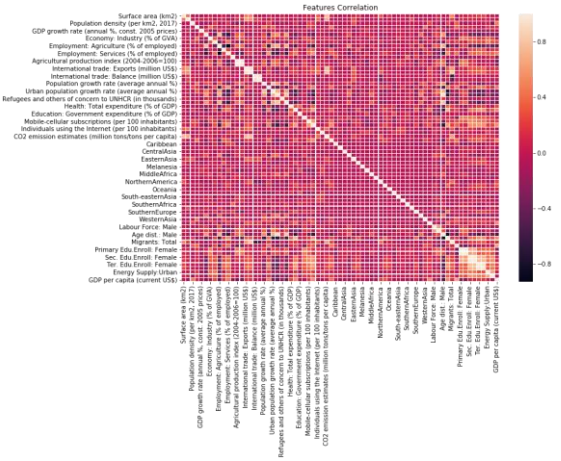
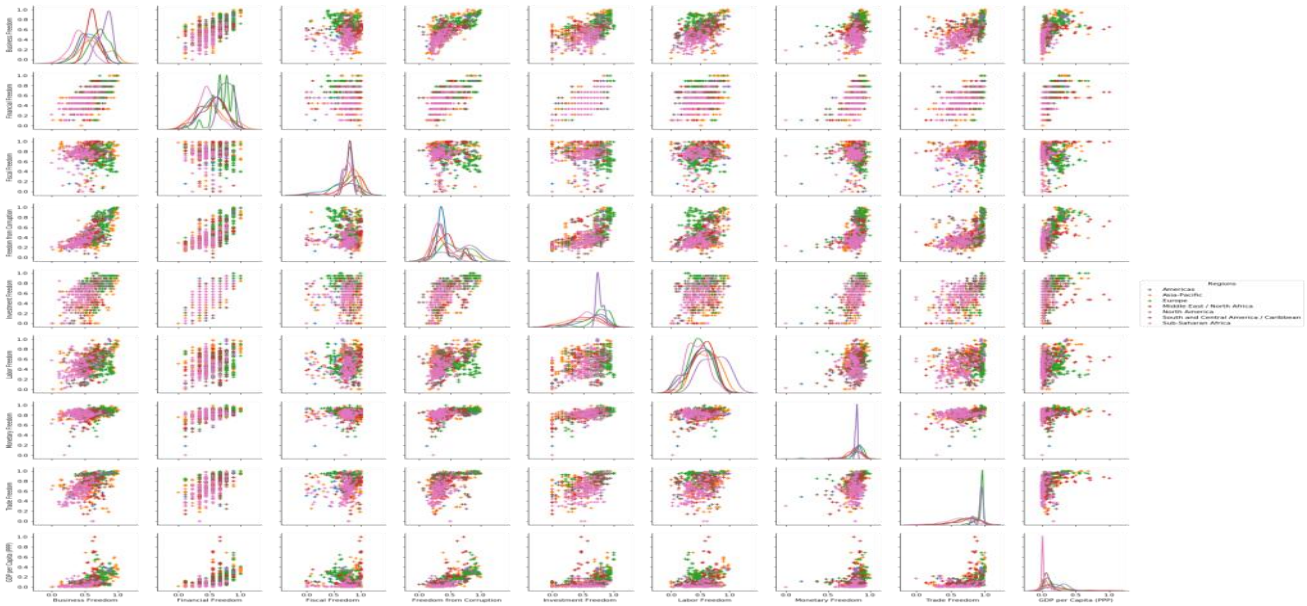


Fig. 3



Algorithms and Techniques

The models used were Random Forest, Decision Tree and AdaBoost regressors, choose for it is due the strong resistance at outliers and easy visualization. The final model is a Decision Tree, decisions trees are models that only uses conditional control statements, simplifying the algorithm create a tree where questions about the features are made if the answer is positive the analysis go for the next branch of the tree and if negative for other way in the tree, in the final branch of the tree there are leafs, for each one a answer. This trees are grow splitting the data point, if there are a number high or equal of leafs (this is how point are called in decision tree algorithms) predefined for a question made a new branch is created, new branch in the tree keep been created until the number of level reach the number predefined or if any are defined until no split be possible cause there are none sufficient leafs to generate a new branch.

The benchmark model use Random Forest, random forest and adaboost are ensemble methods, but different class of ensemble, random forest is a bagging method and adaboost a boosting method. Bagging methods give the same weight for each of the weak learners that compose the strong one.

Many models are tried to achieve the goal some techniques were tried:

- PCA
- K-best Features
- GridSearchCV
- Scaling

The whole techniques use scikit learn algorithms, prepossessing and data cleaning were made with pandas. The evaluations of models were made by a function written to calculate MAPE and MSPE. The train of the regressor used were made splitting the data set in two a third split for evaluation are not used because the Grid Search of scikit learn library has a k fold algorithm built in by standard.

Benchmark

To create a initial benchmark for the this project, a Random Forest out of the box were used, using the kaggle benchmark and no feature selection, scaling or transformation. A function that calculate the MAPE and MSPE were used to evaluate this model and how well it can predict GDP per capita, some prepossessing were needed, some features in the dataset were not compiled by UN for processing a dummy transformation were needed and some columns in it became two features, so the actual benchmark model has 70 features.

Methodology

Data Preprocessing

The processing of the data was made as follows:

- Data acquisition, because there are many missing data in the kaggle data set.
- Data cleaning – data point with missing features were removed.
- Data transformation – some columns in both dataset were not number but text and some were transformed in two columns.
- Data transformation – In both dataset the region features got the same treatment, it was transform in columns for each unique region.
- The data were divided into a training set and a validation set called test.
- Data Selection – Models with k best features were evaluated.
- Data Transformation – Models with PCA were evaluated as well.
- Data Transformation – The final model uses scaled features.

Implementation

The implementation was made in three steps:

- Models training
- Models scoring
- Tuning

All steps were done using Jupyter Notebook titled 'kernel'. In the first step some different models were trained, all of them based in tree even the ensemble methods used use a random forest.

This first step can be divided in:

- Loading the data
- Preprocessing (Section Data Preprocessing)
- Splitting in train and test set – 30% of the data were used as test set.
- Using models out of the box (no tuning)

The second step was evaluating make a comparative measurement model against model, a function was written to do so, this function compute the mean absolute percentage error and the root mean squared percentage error. The scores were compared using tables and graph as show in Fig. 4 and 5. This function had to be written more than one time, at first I decided to do not use any function but re write the code but write a function proved to be better, the particular scoring function was a challenge it include more knowledge in pandas library that I at first had, the the fundamental change in it is include an algorithm to do multiple runs and calculate the average, it was found that cause the size of the dataset that is really small a bad split really affect how each model been scored act, to solve that a multiple run a a score of the mean of single runs was written, in the notebook one can change a parameter an increase the number of runs and with that make the average of the multiples runs be almost the same, and that is actually done.

Fig. 4

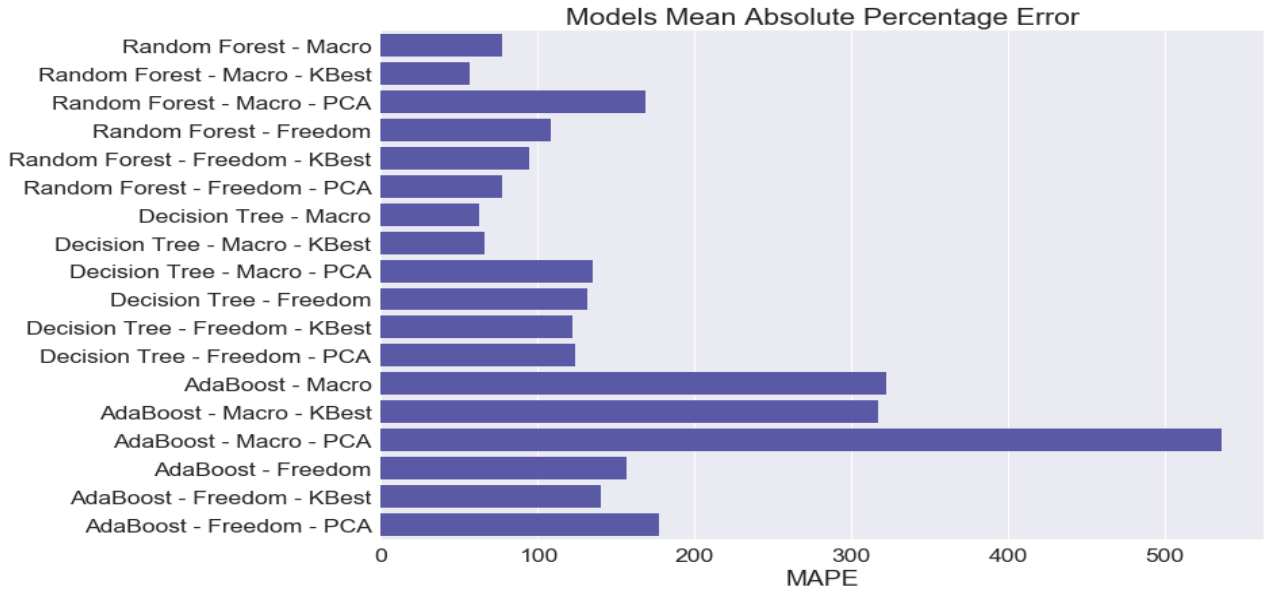
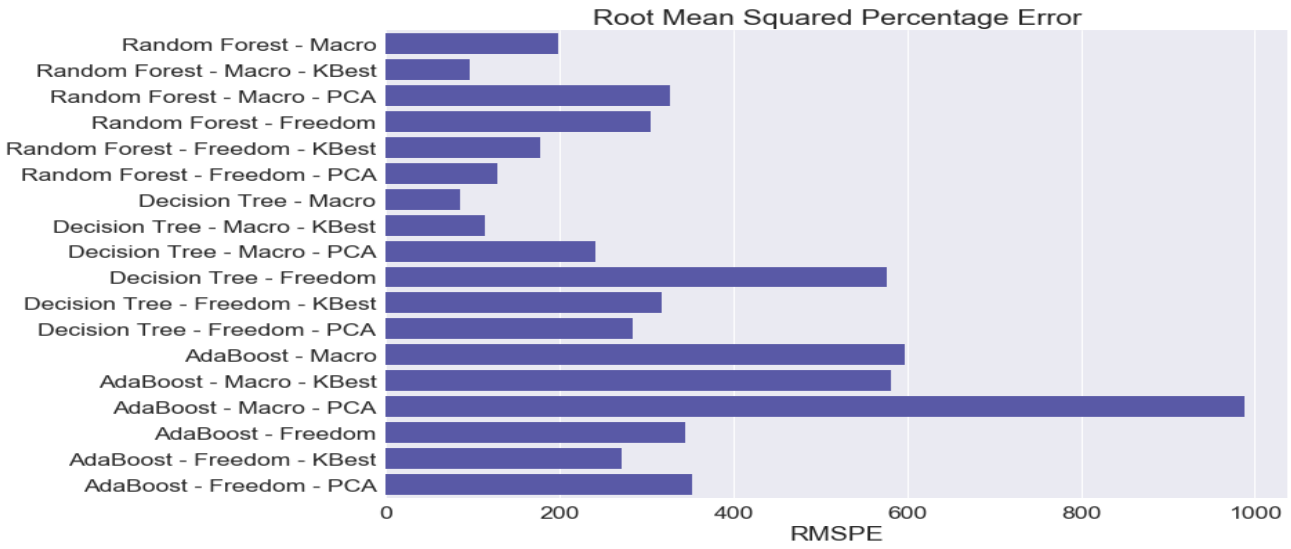


Fig. 5



For the tuning of the hyper-parameters was used grid search, the scikit learn GridSearchCV. A small subset of these parameters can make a huge improvement on predictive or computational performance of the model while others can be left of theirs default values [1]

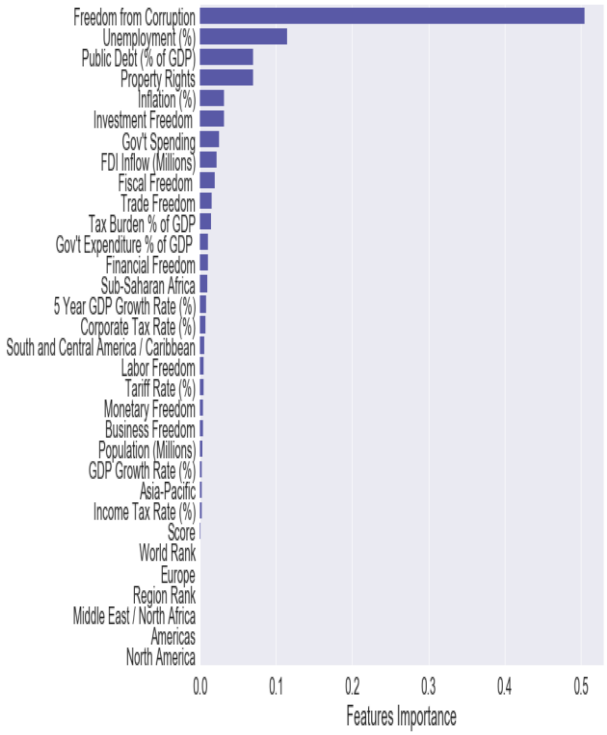
[1] http://scikit-learn.org/dev/modules/grid_search.html#grid-search

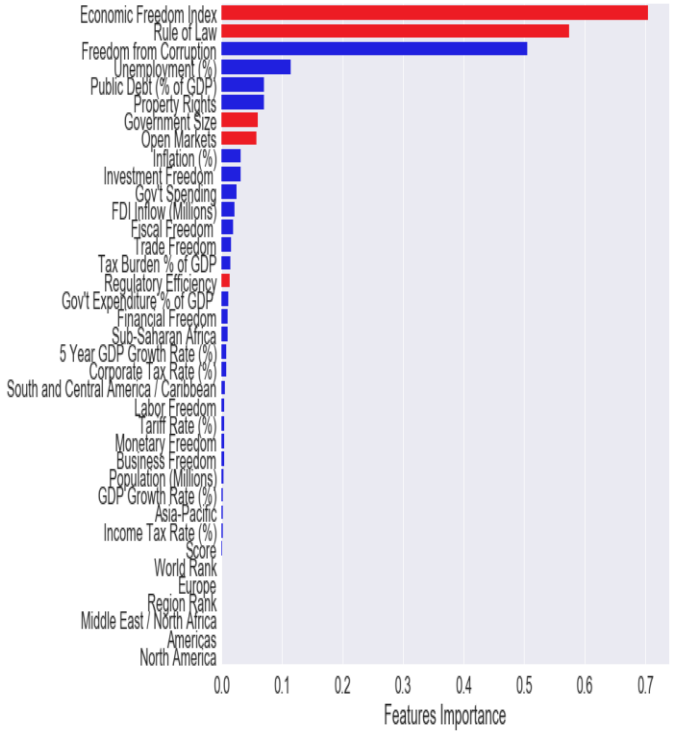
The data for the exhaustive search was not divided because the GridSearchCV uses a cross validation technique k-fold with 3 fold as default number of folds, but to score the models was used the set separated before.
The final model is a decision tree with tuned hyper-parameters as follow:

- Max Depth: None
- Min Samples Split : 2

With this the model do only fifteen question even having more than double of it in number of features, more depth in decision trees models can cause over fitting and make the model act poorly in the test set.

The features importance in this model is shown below:

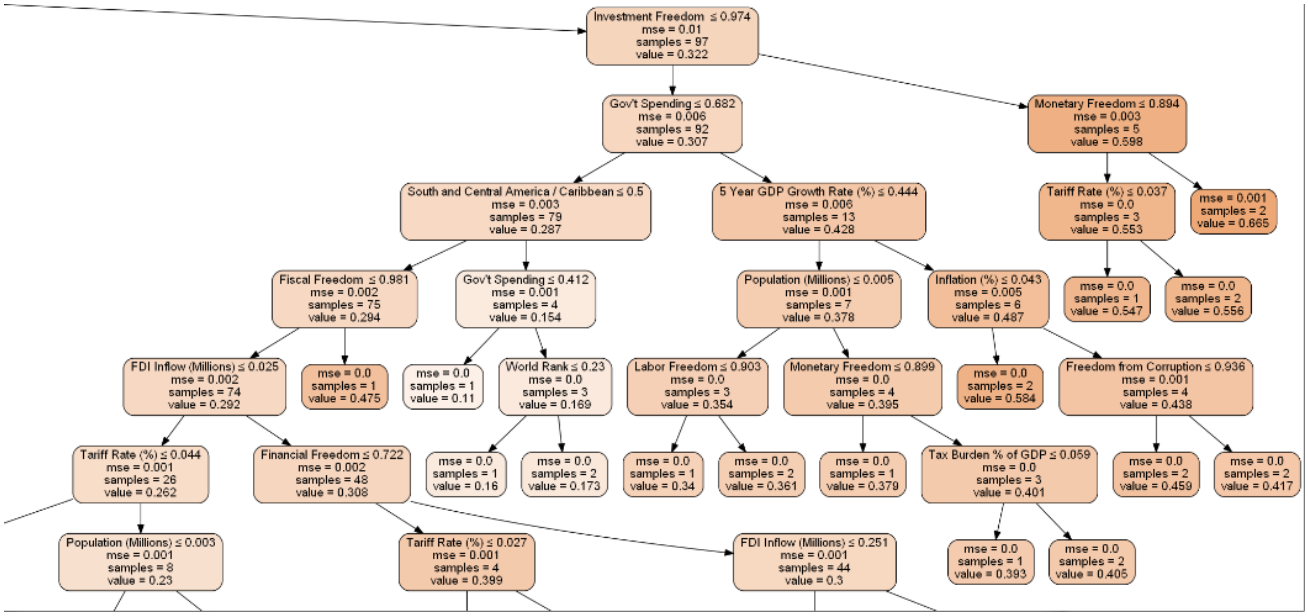




Bars in red are not actual features are groups of features made to have idea of how much the freedom index is important in the model. In the fig. 8 there is the tree used by the model the fig. 9 is a zoom in part of it.

Fig. 8

Fig. 9



Refinement

The benchmark model used achieves 76.96 and 198.33 in MAPE and RMSPE respectively. The benchmark model uses the kaggle data set to predict GDP per capita and a random forest algorithm out the box to do it. Others models was tested as show in fig. 4 and fig. 5 not only different models were used but features engineering techniques.

In this work one can perceive that PCA and K-best can really help but not always one technique will be the best. K-best got better results in terms of MAPE and RMSPE than PCA for the kaggle dataset using the random forest algorithm, but using the Heritage one the opposite happened and using a decision tree instead of a random forest both made the model act poorly.

The initial refinement was just search for the best model; the best model uses the kaggle dataset with K-best features where the best 10 features where used to score 56.85 MAPE and 96.7 RMSPE

As there are more data available in the heritage web site it was loaded and transformed in a single set, using data from 2013 to 2018, 935 data point after cleaning became in 771 data point. And with it using a decision tree model it improves to 44.09 MAPE and 110 RMSPE. With tuning it got to 33.02 MAPE and 65.278 RMSPE.

It was found that for this compiled dataset decision trees out of the box act better than random forest used with no tuning, it only got 59.47 in MAPE and 147 RMSPE that is worse than the decision tree but with tuning the random forest model got 14.947 MAPE and 28.04 RMSPE, it is a improvement of four times MAPE and five times in RMSPE only with tuning. It was found that for its data set and model features engineering make the model act worse, the techniques used were feature selecting (k-best), feature transformation (PCA and scaling), the random forest was not affected well for scaling but the decision tree algorithm was, achieving 0.5 MAPE and 2.42 RMPE in the best run of the notebook.

Results

Model Evaluation and Validation

During the project a validation set was used to evaluate the model and a random forest used as a black box model with the kaggle dataset as benchmark.
The final model was chosen for it performance that is 153 and 92 times better than the benchmark and scores 262 and 238 times better than the first decision tree used in terms of MAPE and RMSPE.

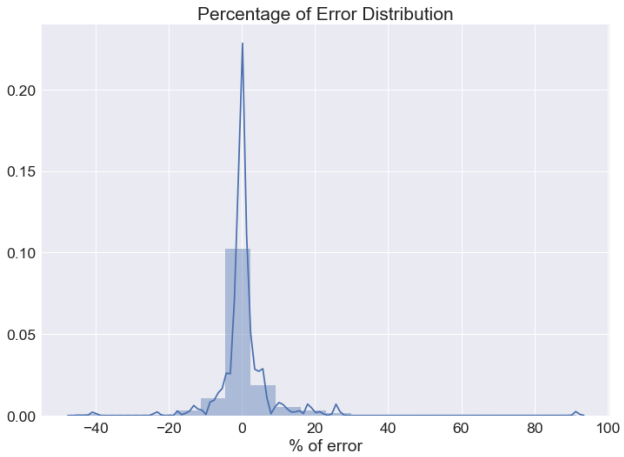
This models in question uses data from 2013 to 2018 and the initial decision tree model used only the 2018, the final model uses the data from 2017 and 2018 but do not uses the feature judicial effectiveness that is the more important feature in the initial models. Beside that increase of 4 times in usable data the model made a performance increase of 5.2 in RMSPE and 80% better of the benchmark that uses the kaggle set. Using 3 times more data point and the half of the number of features the final model performance is 100 times better in terms of MAPE and 40 times better in term of RMSPE than the best model using the data set of kaggle.

All this number are calculated using the best run of the notebook created, the performance change quite a bit in run to run due the hyper-parameters search made by GridSearchCV that uses k-fold cross validation technique, different result in it is explained by the randomness in the split, the worse split cause a performance of 55 MAPE but in the mean the model performs better than 5% in MAPE and 15 % in RMSPE.

Justification

To comprehend how well the model perform only the score are not sufficient, understand how error is distributed is vital, this model perform fine but for some point error is still large, going to 100 in best run to 200 % in a run where it achieves 5 of MAPE, the best run got 0.5 MAPE ten times better. Besides that the actual models show robustness in it results, having an overall achievement ten times worse it got only double of the maximum error. That is significant because it shows that the actual model really learns about GPD per capita and how the features affect it.

Using Tkinter library GUI made in the notebook one can uses the model prediction to understand the composition of GPD per capita in this is putted sections of sliders that the user can change and see the prediction, with that the users can better understand each feature and how it change the prediction, this model performance is good but is not suitable for this class of application were the user change value and with minor change is expected a new prediction. For this a linear models will be better suitable but the performance of this kind of model for this task is really bad. A neural net may give a better



application were minor change cause change in the output of the model and with that the user can understand better GPD per capita and which feature has to be improved to give maximum output (see the Improvement section)

Conclusion

Free-form Visualization

Some minor change in the slide can cause huge difference in the GDP per Capita predicted and some huge change any dollar in the prediction at all. Besides that fact over the final model the utilization of a graph interface show that region is not a big deal and talking about GDP, but corruption is actually is, and in fact is the first thing that the model considers as shown bellow in fig.12

The graph interface was made to interact with the model, not all features is present in the interface this features are calculated, using the features present in this interface as presented in the justification section the interface are not very responsible and it is due the fact that the model uses not a formula where each feature has a weight but a decision tree and because that value that do not moves into another branch of the tree get the same prediction. So the better uses the interface some knowledge of the decision tree helps not get stuck, the initial branch of the tree can be seen in fig. 12

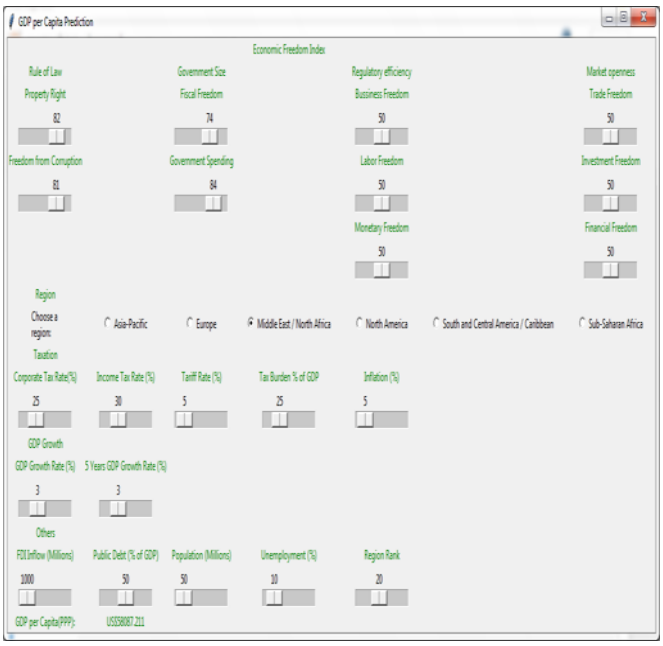
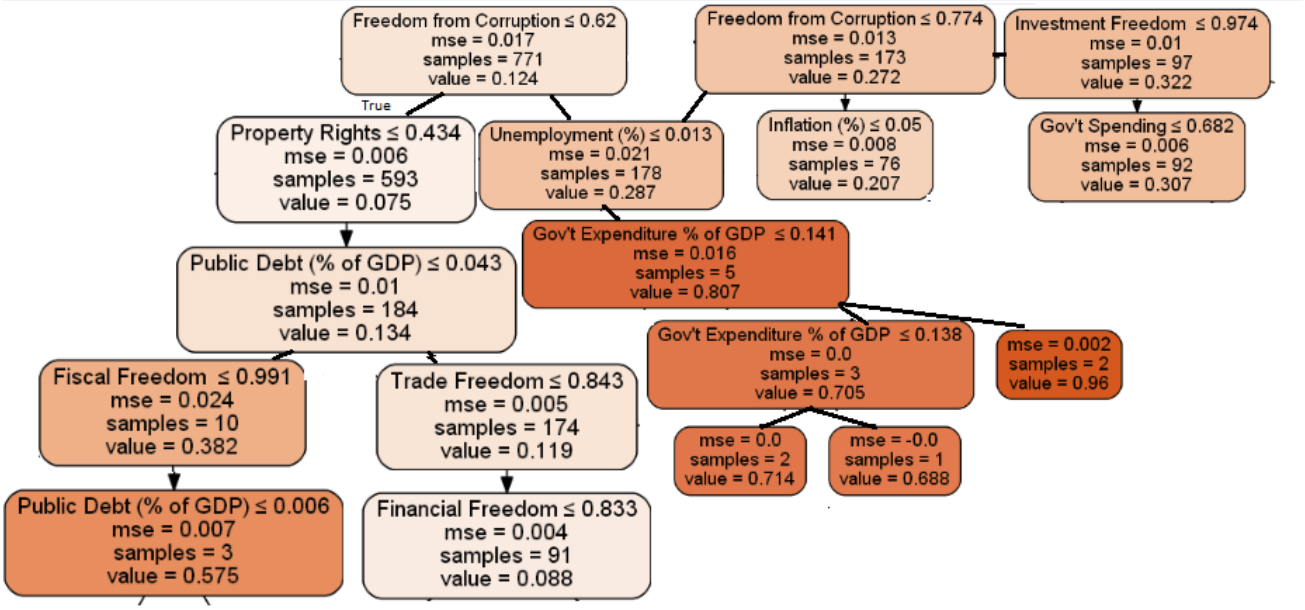


Fig. 12



Reflection

This project follows some steps that were not at all always linear but in resume it follows this:

1. Relevant public data set for the problem were found (see Analysis section)
2. The data were loaded and prepossessed (see Methodology section)
3. A benchmark was created
4. The regressor was trained using the prepossessed data (different features engineering and models were tried to find better one and the better set of parameters)
5. A Tkinter interface was made

The fourth and fifth step, the last one of this project were the more difficult, the library Tkinter was near used before by me. It took me some time to figure out how to make what a want and it is not quite beautiful interface but it works as I wanted. Some improvement in design can be made and the lack of predictions for values that lies between branches should be fixed.

The fourth step push me up to use the knowledge collect in features engineering, the understand of the characteristics of the dataset was fundamental the select the right techniques, PCA and features reduction do not helped cause the dataset used has few features (but helped with the kaggle data set with double of number in features). PCA do not either this data set has high amount of outliers and variability

Improvement

To achieve a responsive interface where minor change in the features led to different prediction a deep neural network probably should be implemented, different solution in neural network can be addressed even CNN that are commonly used for image classification can be used for the task, one can transform the time series data into graphs and uses this graph as input, and doing so the lack of data for some years in some countries become information too, to achieve a tool that can actually be used by government only one target is not sufficient, each one of the features in the heritage has some correlation with the others so a improvement in one probably cause change in the others and in this project it is not taken into account. To do that will be need an architecture having two layers of models in the first layer models that using the feature in change can predict the values of all features and in the second layer a model like the final model of this project, which takes all features and predict a target variable.