Machine Learning Engineer Nanodegree

Capstone Proposal

Isac Soares Camara Junior

August 9st, 2018

Proposal

Predict GDP per capita by econometrics data, social information and liberty.

Domain Background

The domain is econometrics and social data like: size of the country, region, population density, how the economy is distributed, how labor force is employ, life expectancy, how much is expend in health, among others ones.given by UN.

And liberty as economic freedom given by heritage index, the world wide know index of economic freedom, that maps: property rights, government integrity, judicial effectiveness, tax burden, government spending, fiscal health and freedom of business, labor, monetary, trade, investment and financial.

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_Predicting_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

Datasets and Inputs

The Dataset used will be a kaggle dataset of information of countries and the economic freedom index by heritage.org.

https://www.kaggle.com/sudalairajkumar/undata-country-profiles

https://www.heritage.org/index/explore

The dataset provided by kaggle is not complete and have some errors, to solve that another fonts will be used such as World Bank, CIA World FactBook and countries government papers.In this dataset are 50 features one that is GDP per capita will be used as target variable of the supervised learning model GDP will be dropped because the correlation with the target variable and population that will be used, 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

The dataset of Heritage have 13 useful features and 188 countries, all continuous data .

This features will be used because it can lead to clues of how is the environment of the given country, knowing that one may predict how the inhabitants of this lives, and how well they perform in economy.

Solution Statement

Find the correlation between GDP per capita and the features for the two datasets, and compare it. build two models that uses of all of the features in the given dataset and measure how well it can make predictions and compare the two models and by that find if is freedom that leads to economic health or is State investment and programs.The goal is find with supervised learning using the features in the datasets as input and the GDP per capita as the target variable which features

or which set of features affect the most the predictions to be able to come with a better model.

Benchmark Model

An out-of-box random forest model will be train as a benchmark model as it was in the paper cited in domain background section. As input for this model only de kaggle dataset will be used, doing that one can compare how well freedom alone can explain GDP per capita, the results of the paper will be used for benchmark as well .

Evaluation Metrics

The models will generate predictions of GDP per capita, the error between the actual value of it and the predicted as percentage of the actual data will be used. The mean square error of this percentages will be used for final evaluation. The error between the prediction and actual value cannot be used directly because the high difference of GDP per capita among countries and the total of countries in the two datasets are different.

Project Design

The Design of workflow will be the data acquisition, some data are wrong in the kaggle dataset and for many countries there are missing data, some research in open data set like CIA World FactBook, World Bank and Country government papers will be used to fill this gaps.

Data cleaning: some investigation will be done to see if there are wrong data, like typos making the decimal point be in the wrong place making a data point(country) larger, smaller, having a really large imports with almost zero exports and etc. To do that I will look for outliers and see if there are any error in the data.

Data Analysis: Heat maps analyses using Seaborn is really simples, one line of code and the work is done, some data analyse for the most correlated features will be done, some discussion of why may this features have a correlation will be given. And further analysis of features correlation and possible feature dropping to make a better model will be presented,

Building Model and Training:The machine learning predictions will be made by decision tree algorithms and ensemble methods: decision tree regressor , random forest regressor, and adaboost and XGBoost.

The benchmark model will be trained with all features of the kaggle dataset and will be a random forest out-of-the-box with no tuning.

A grid search will be made to tune the others models and the best one will be used to compare with the benchmark model.

The decision by decisions trees is due the fact that this models can be easily visualized and this is the main goal, create a tool to analyse where to put the money to have better economic health, and what to change to get that. The possibility to have a visual tool to aid in this decision worldwide is a must have.

The ability to compare two different approaches and see how each can affect the population is really interesting for future political discussions.