Data oriented programing - exercise 3: Summary

Group 46: Tin Oroz, Luka Došen, Tomislav Seljan

September 26, 2022

1 Introduction

Goal of this exercise is to explore how many people in the world have access to electricity and how has this changed over time. We also want to check if there are correlations between accessible electricity and some of the potentially correlated statistics. Metrics for which we are going to check correlation are: Literacy rate, GDP per capita, Electric power consumption, Life expectancy, Poverty rate.

We also wanted to explore differences in access to electricity in rural and urban areas. From this exploration, we are going to try and build a model, which estimates a percentage number of access to electricity for a country in a given year from all the available data. All of the data sets contain aforementioned statistics for each country per every year between and including 1960-2020.

2 Insights into the data

2.1 Cleaning the data

We had a few problems with our data that needed to be dealt with. First and biggest was the missing values. Data sets had a lot of missing values we had to either remove or try to approximate. If we had just removed all of the rows containing missing values we would have removed too much data which is not good. For replacing missing values we use techniques such as replacing missing values with values from similar columns, predicting them or approximating them by means. One of the biggest removals we did in our data, to make it fit the predictive model, was to remove values before 2000. This had to be done because a lot of third world countries did not have any data before those years. This can be seen in figure 1. This shows that access to electricity was going down from 1990 to 2000. This can be explained because we had more and more data from Third World countries. Since most of our data is in format of a percentage number, for outlier handling we just checked if the numbers are outside the [0, 100] interval which was not the case.

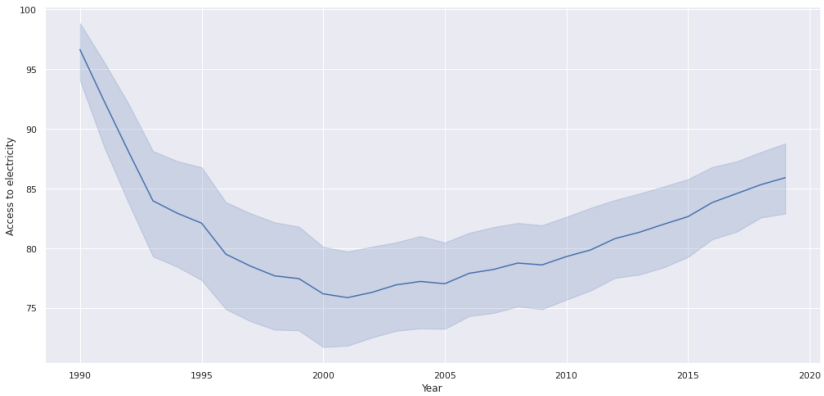
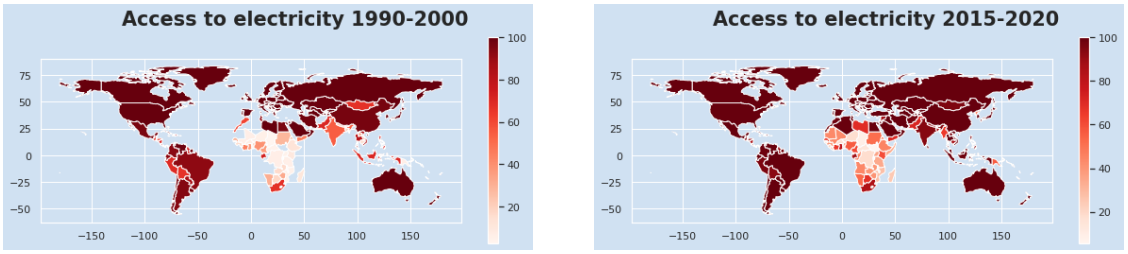


Figure 1: Access to electricity



(a) (b)

Figure 2: Access to electricity in world in different time spans

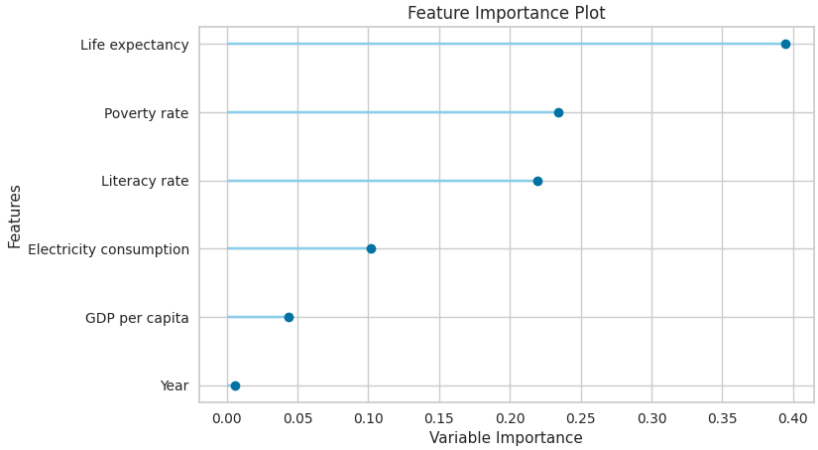


Figure 3: Feature importance

2.2 Access to electricity

Main conclusions we draw from this exploration is that access to electricity over the years changed for the better. The main difference was in the Third World countries. This can be seen from Figure 2. The biggest improvement can be seen in countries in Africa.

Another thing we explored was the influence or correlation of other different metrics we included in our project. To get the best insight into this, firstly, we constructed a correlation matrix. From the matrix the significance of the correlation between Access to electricity and other variables can be seen. Life expectancy and Literacy rate had the highest positive correlation of 0.8. Poverty rate also has a high, but negative, correlation of -0.8.

Another way we wanted to explore this was training a predictive model that predicts access to electricity based on a year and all of the other metrics we included in the project. We explored some potential models and the best fit seemed to be Extra trees regressor. The model had an R2 value of 0.96 which is pretty good. From a trained model we would then analyse feature importance to get a deeper knowledge about correlation between access to electricity and other metrics. This can be seen in Figure 3. From this we can see that the variable Life expectancy had the biggest influence. An interesting finding is that the last two variables ranked by correlation were year and Electric consumption which were expected to have a big influence, but in reality they do not.