DengAI: Predicting Disease Spread

Davide Lagano

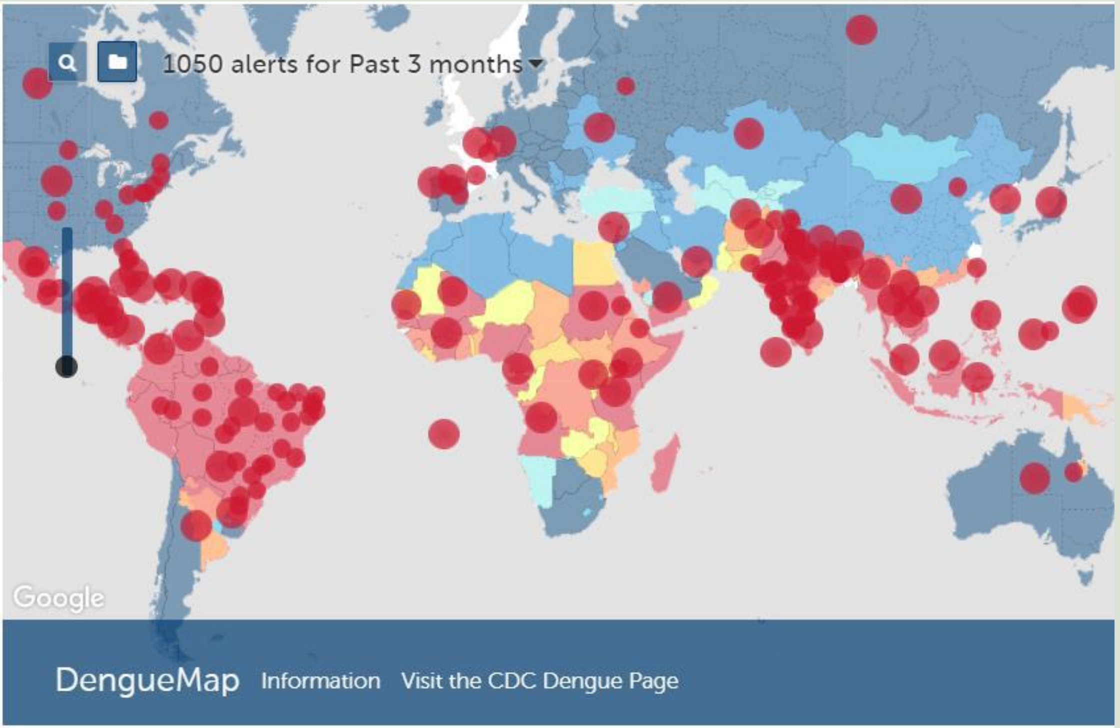
Introduction

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world.

Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation.

In recent years dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half-billion cases per year are occurring in Latin America:

An understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.



Purpose

The goal of the competition is to predict the number of dengue fever cases reported each week in San Juan (Puerto Rico) and Iquitos (Peru) based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

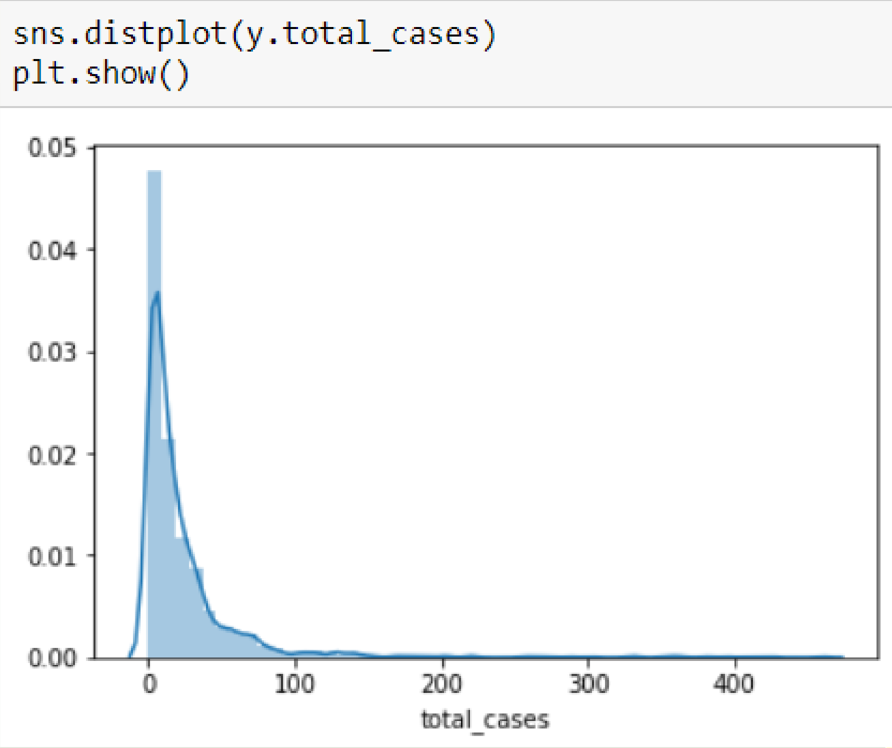
Performance is evaluated according to the mean absolute error.

Method

In this project have been used different pre-processing for each models.

models used:

- Simple linear regression
- XGBoost
- Negative binomial regression



```
sns.distplot(y.total_cases)
plt.show()
```

Data

The features dataset is composed by 1456 row and 24 columns. Available at https://www.drivendata.org/.

Results

- Simple linear regression:

L1: mean absolute error of 8.20

L2: mean absolute error of 7.97

- XGBoost:

mean absolute error equal to 18.27

- Negative binomial regression:

cv error: 17.42

train error: 18.32

Conclusion

For the first model, simple linear regression is better than a random model. Moreover, recursive feature elimination improves the accuracy. Furthermore, regularization helps to the overfitting.

As a future improvement of the paper could be interesting to indagate why, in the cross-validation, the array of the different cities are so different. Probably the improvement of the tuned model is due to the overfitting.

The results of the negative binomial regression are quite nice, it’s impossible to compare this result with the others, but probably this is the model to start with to improve the paper.