DengAI: Predicting Disease Spread

HOSTED BY DRIVENDATA

Introduction

It can be seen that the transmission of Dengue fever is related to various climate variables such as temperature and precipitation. Although the relationship between the transmission of Dengue with climate variables is complex, many experts argue that climate change is likely to significantly affect the spread of Dengue fever.

The problem we are trying to solve is, to predict the number of dengue cases (number of patients who are caught with Dengue) each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation, etc.

The competition provided a data set containing around 1450 entries, of climate conditions and reported dengue cases in two cities; San Juan and Iquitos. The goal is to predict the total_cases label for each (city, year, weekofyear) in the test set.

Methodology

First, I preprocessed the available training and testing data. I split the dataset based on the city. To fill in the missing values, I took several approaches. One approach was to fill the missing values by considering highly correlated redundant features (for example station_avg_temp_c and avg_air_temp_k). After experimentation, I recognized that filling the missing values with a mean value of the differences would improve the results rather than using options like forward fill, backward fill, or ignoring the missing values. As a normalization method, I transformed each value into a z score. The reason behind normalization is to make training data less sensitive to the scale.

The next requirement was to do feature engineering. Rather than choosing only the features which are highly correlated with the total number of cases, I plotted the graphs for each feature value including total cases. Here, I identified hidden patterns between features as well. One such pattern I identified was features like dew point temperature, specific humidity, min/max/average temperature peak every 52 weeks or around 52 weeks similar to the total number of cases in San Juan. Though that is the case, a direct correlation was not shown as the quantitative change of these features are not proportional to the total number of cases. Therefore, in feature selection, I considered correlation as well as hidden patterns between features. I created new features using moving window average (window size = 52) for features like precipitation to capture the trend over that period of time.

Furthermore, I had to select a suitable machine-learning algorithm and evaluate the results. For evaluation, I split the dataset as training and cross-validation data set randomly. First, I trained are tested on several machine learning models; linear regression, support vector regression (SVR), gradient boosting, and random forest regressor. I got better results from gradient boosting and random forest regression for San Juan. Linear Regression, SVR and gradient boosting performed better for the city Iquitos. I selected those models for further processing. Next, I used GridSearchCV to tune the hyperparameters of each of those selected models.

Finally, I ended up with Gradient Boosting model for San-Juan and SVR with RBF (radial bias function) kernel for Iquitos.

Results and Analysis

Results obtained using different regression models are shown in Table 1.

Model	MAE for SJ	MAE for IQ
Linear Regression	27.826	6.334
K-Nearest Neighbour Regression	25.240	6.033
Support Vector Regression (linear kernel)	23.035	5.585
Support Vector Regression (rbf kernel)	21.931	5.551
Gradient Boosting	20.644	5.523
Random Forest Regression	18.984	5.441

Table 1 shows the mean absolute error for the cross-validation set I split. I have not used a neural network model because the available data set is less than 1500 records. We can observe that the linear regression fails. One possible reason might be that the correlation between each individual feature and the total_cases was less. Therefore we can say that the relationship between those features and total_cases is nonlinear. Another possible reason might be the correlation between selected features. We can observe that the error in K- Nearest Neighbour regression algorithm is also high because of the high dimensionality. Gradient Boosting and Random Forest Regression models perform ahead of other models possibly due to the ability to learn non-linear relationships and robustness to outliers. Surprisingly Support Vector Regression with a ‘linear’ kernel performed better for Iquitos. That observation might be because the Support Vector Regression model does not change its model parameters unless the threshold is met. By

observing the above results, I proceeded with Random Forest Regressor and Gradient Boosting Models for San Juan and Random Forest, Support Vector Regression and Gradient Boosting model for Iquitos and tested the results with hyperparameter tuning using GridSearchCV. The winner for San Juan was the Gradient Boosting model with a mean absolute error of 16.285. For Iquitos, the best results were achieved by further tuning the Support Vector Regression model having the ‘rbf’ kernel, with a mean absolute error of 5.043.