NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

BS6200

DengAI: Predicting Disease Spread

Members:

ZHANG CHENXI

Content

I. Introduction

Dengue fever is a mosquito-borne tropical disease caused by the dengue virus. Symptoms typically begin 3 to 14 days after infection. These may include a high fever, headache, vomiting, muscle and joint pains, and a characteristic skin rash. The earliest descriptions of an outbreak date from 1779. Its viral cause and spread were understood by the early 20th century. Dengue has become a global problem since World War II, and the reported had increased significantly worldwide. In 2019, a significant increase in the number of cases was seen. It is common in more than 120 countries, mainly in Southeast Asia, South Asia, and South America. About 390 million people are infected a year, about half a million require hospital admission, and approximately 40 thousand die. Apart from eliminating the mosquitos, work is ongoing for medication targeted directly at the virus. It is classified as a neglected tropical disease. Since it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide. Thus, an understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.

II. Problem Statement & Objective

This project will use machine learning techniques and regression models to predict the number of dengue fever cases reported each week in San Juan and Iquitos. The host of this project has provided 3 datasets, which include the dengue train feature set, the dengue train label set, and the dengue feature test set. All of the sets are scaled and ordered in chronological order. The goal is the minimize the mean absolute error between the predicted result and the actual result.

III. Dataset

A. Data Introduction

The data files consist of the dengue train feature set, the dengue train label set, and the dengue feature test set. The feature set includes five different categories.

City and date

- “city”: the City abbreviations, where“sj” for San Juan and “iq” for Iquitos,
- “year”: the year when the case was recorded
- “week_start_date” represents the week when the case was recorded

Daily Climate Data

This part includes the daily climate data measured by the weather station

- station_max_temp_c: Maximum temperature
- station_min_temp_c: Minimum temperature
- station_avg_temp_c: Average temperature
- station_precip_mm: Total precipitation in millimeters
- station_diur_temp_rng_c: the Diurnal temperature range

Climate Forecast System Reanalysis

This part includes the climate forecast focusing on air temperature and humidity.

- reanalysis_sat_precip_amt_mm: Total precipitation in millimeters
- reanalysis_dew_point_temp_k: Mean dew point temperature
- reanalysis_air_temp_k: Mean air temperature
- reanalysis_relative_humidity_percent: Mean relative humidity
- reanalysis_specific_humidity_g_per_kg: Mean specific humidity
- reanalysis_precip_amt_kg_per_m2: Total precipitation in kilogram per square meter
- reanalysis_max_air_temp_k: Maximum air temperature

- reanalysis_min_air_temp_k: Minimum air temperature
- reanalysis_avg_temp_k: Average air temperature
- reanalysis_tdtr_k: Diurnal temperature range

Normalized Difference Vegetation Index

The part of the data is a simple graphical indicator that can be used to analyze remote sensing measurements, often from a space platform, assessing whether or not the target being observed contains live green vegetation.

- ndvi_se: Pixel of the southeast of city centroid
- ndvi_sw: Pixel of the southwest of city centroid
- ndvi_ne: Pixel of the northeast of city centroid
- ndvi_nw: Pixel of the northwest of city centroid

B. Data Analysis and Visualization

For analyzing the feature set, I combined the features both from training and test data sets. Since the datasets include the features of two cities, I decide to separate them and do analyses on each one.

Missing Values

The analysis shows that besides the city and date parameters, all other features have missing data. Since the rest variables are continuous, thus filling them manually is not an option; and the proportion of missing values is quite big, thus ignoring them is not a good choice as well.

In the remaining options, I checked the correlation between each variable. Figure 1 is the correlation heatmap between each variable. From this figure, it is easy to observe that there exist some high correlations between several features, also, it is reasonable to believe that some variables are related to each other, such as the Maximum air temperature, the Minimum air temperature, the average air temperature. Thus instead of filling the missing values with column mean, I decide to use linear regression to interpolate the missing values. In addition to this, it is good to remove high correlated features in supervised learning, thus I decided to drop the terms that correlate greater than 0.95 or smaller than -0.95. The reason to do so is to prevent muticolineariy, which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy, so that affects calculations regarding individual predictors As a result, the features “reanalysis_avg_temp_k”, “reanalysis_sat_precip_amt_mm”, “reanalysis_specific_humidity_g_per_kg” can be eliminated since they have very similar features left in the dataset.

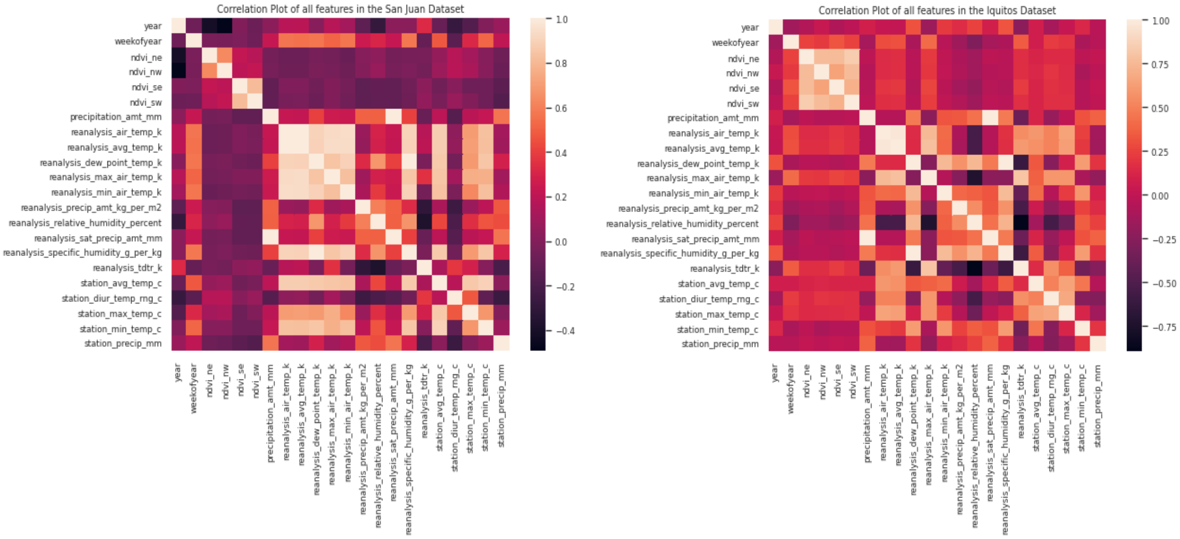

Figure 1. Pearson’s correlation coefficients heatmaps

Outliers

Figure 2 shows the boxplot, for observing the distribution as well as the outliers of each feature. I decided to ignore the ‘year’, ‘Week_start_date’, and ‘Weekofyear’ for now and judge solely on the continuous features. From the boxplots, it is easy to observe that most of the data points from the features from the Climate Forecast System Reanalysis measurements are deviate from their quantiles, which are outliers. First of all, I noticed some temperatures’ features are measured in Kalvin and some are measured in celsius. So I converted all Kalvin degrees to Celsius degrees. Then to Center and scale, I applied Z-score normalization to the feature sets by removing the mean and scaling to unit variance.

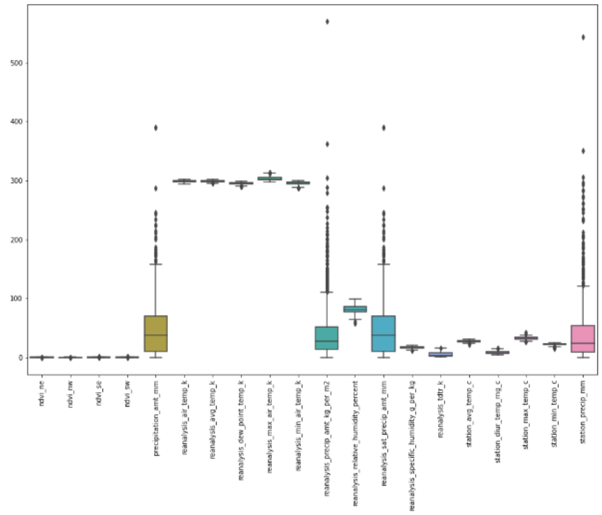

Figure 2. Box Plot

Feature Selection

Figure 3 and figure 4 show the correlation of between each feature and the total cases in each city in ascending order, which represents the feature importance of each feature. From each plot, we can observe that "reanalysis_dew_point_temp_k", "station_avg_temp_c", "reanalysis_max_air_temp_k", "station_max_temp_c", "reanalysis_min_air_temp_k", and "reanalysis_air_temp_k" are the features with highest importance in San Juan, and "reanalysis_dew_point_temp_k", "station_min_temp_c", and "reanalysis_min_air_temp_k" are the features with highest importance in Iquitos.

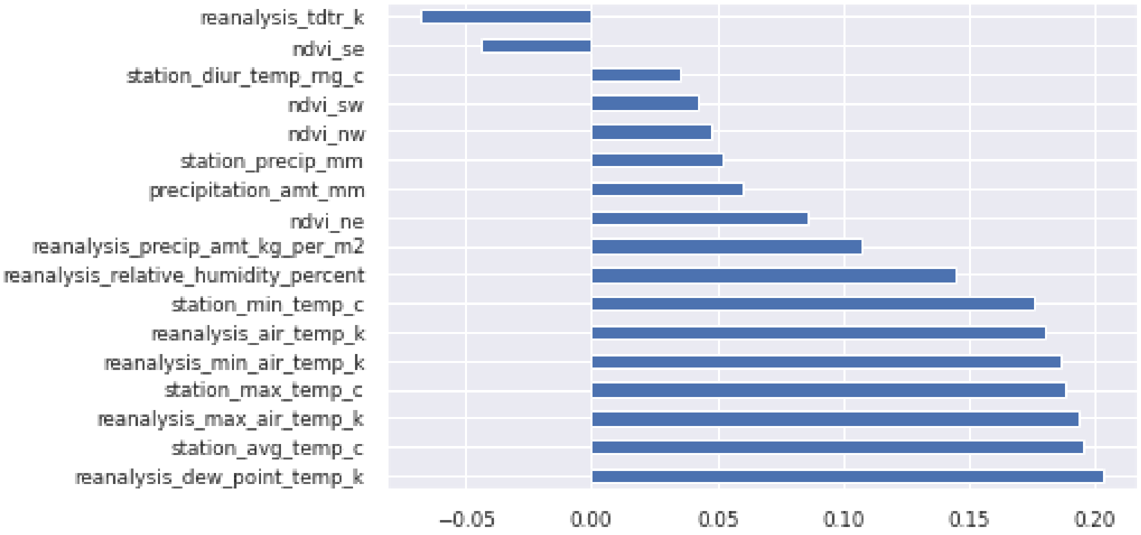

Figure 3. Feature Importance of San Juan

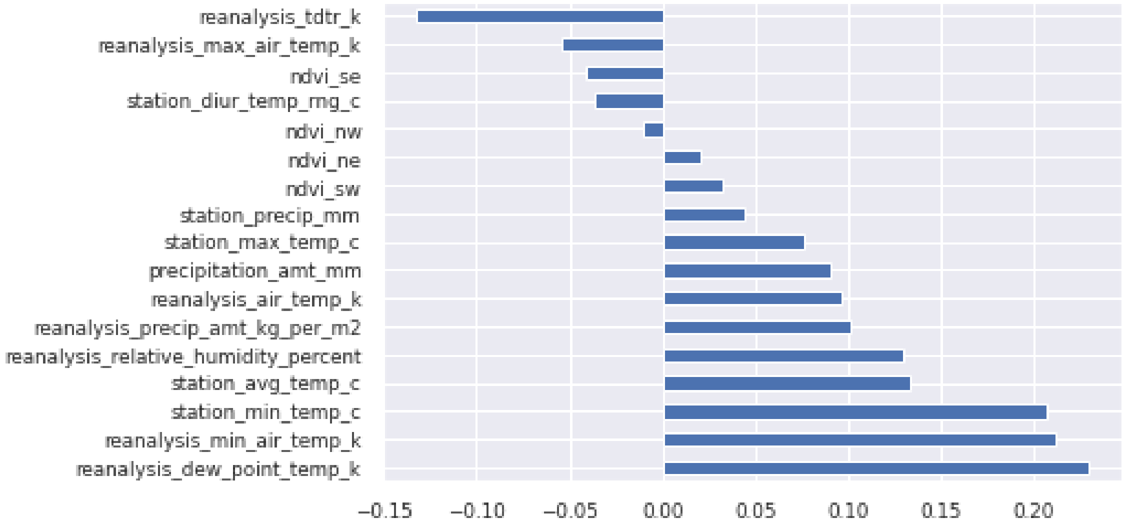

Figure 4. Feature Importance of Iquitos

In addition, I used the time variable to see the hidden periodic trend of each feature. Rather than choosing only the features which are highly correlated with the total number of cases, I plotted the seasonal decompose graphs for each feature value including total cases. Figure 5 shows the selected features' plots that shown a significant periodic pattern in time. One such pattern we can see that features like dew point temperature, specific humidity, min/max/average temperature peak every 52 weeks or around 52 weeks are similar to the total number of cases in San Juan.

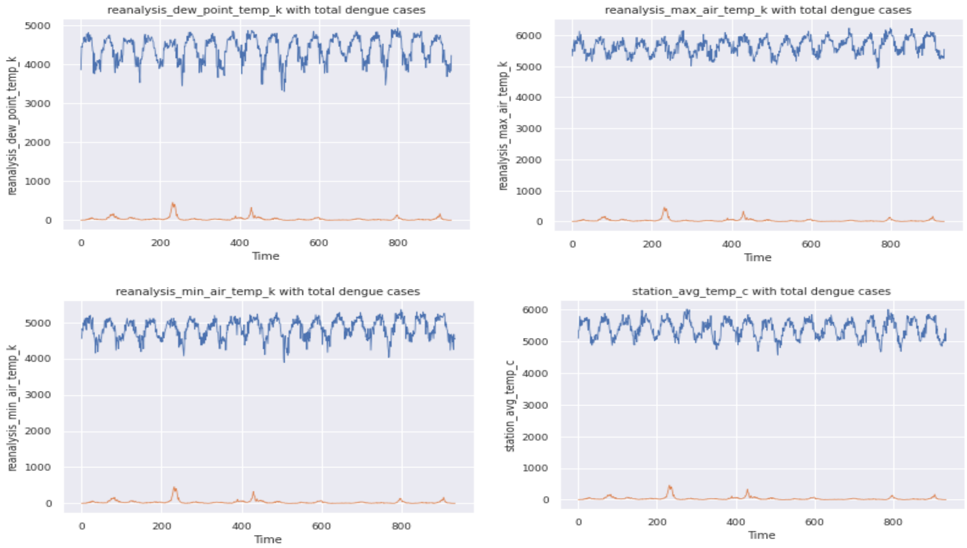

Figure 5. Selected Seasonal Decomposition

After cross compare the correlation with the periodic pattern, for San Juan, I decided to keep "reanalysis_dew_point_temp_k", "station_avg_temp_c", "reanalysis_max_air_temp_k", "station_max_temp_c", "reanalysis_min_air_temp_k", and "reanalysis_air_temp_k"; similarly, I decided to keep "reanalysis_dew_point_temp_k", "station_min_temp_c", and "reanalysis_min_air_temp_k".

C. Model Selection

Figure 6 shows the histogram of the total cases in order to see the distribution of the target. From the plot, we can see that the distribution is skewed to the right. The shape is very similar to the Poisson distribution and negative binomial distribution model.

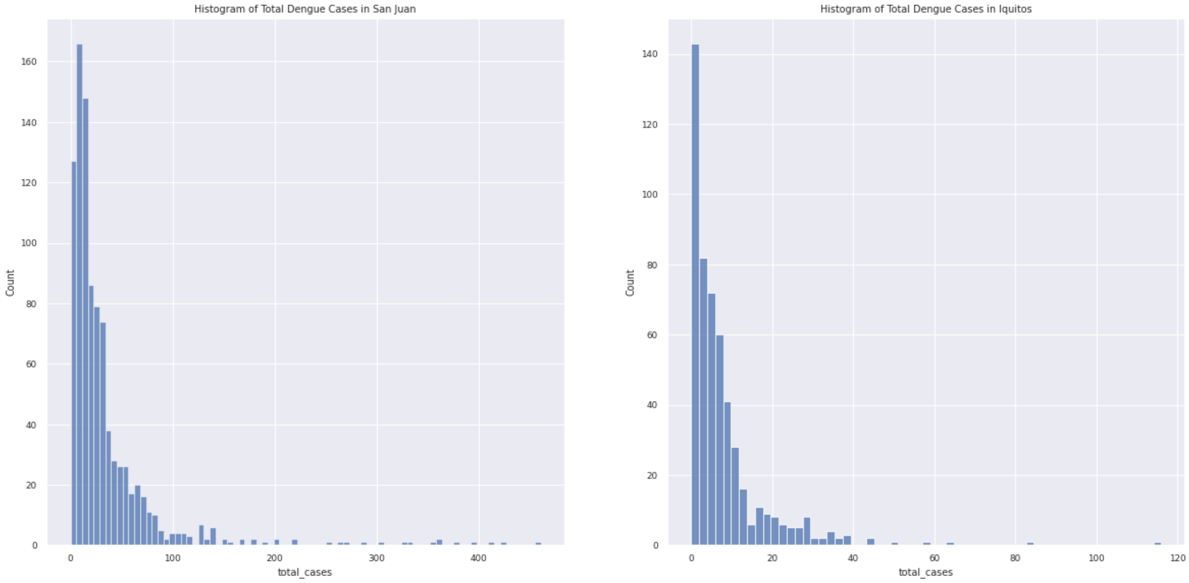

Figure 6. Histogram of the total cases

As a result, my candidate models include negative binomial regression and Poisson regression model with the selected features, with linear regression as the benchmark. In addition, I also include Random Forest Regression to observe the mode ensemble result. Unlike the regular Random Forst, I assigned a weight to each feature based on their correlation to boost the result.

IV. Machine Learning Models

A. Possion Regression

Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables, where it assumes the response variable has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

B. Negative Binomial Regression

Negative binomial regression is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model, which is based on the Poisson-gamma mixture distribution. This formulation is popular because it allows the modelling of Poisson heterogeneity using a gamma distribution.

C. Support Vector Machines

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression, which combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. It usually performs great on many problems, including features with non-linear relationships. In addition, I added weights assigning to the model, which is based on the feature's correlation. So I extract the correlation feature map from the features, and normalized them to percentages and multiply with the features. This grantees that high correlation features gets higher weights, and low correlation features get low weights.

V. Traing & Testing

In the project, the validation technique being used is called leave-one-out cross-validation. Compare to k-fold cross-validation, Leave-one-out cross-validation is approximately unbiased, because the difference in size between the training set used in each fold and the entire dataset is only a single pattern. In addition, it is computationally inexpensive for models like regression, so it is suitable for imputations like MAE or MSE.

The overall Training and Testing Procedure are the following steps:

1. Use each model to fit the training set and validation set, with cross-validation to find the best hyperparameter for the model which minimizes the mean absolute error
2. Apply the model to the test set to see the test MAE and use it to compare with other models

VI. Results & Comparison

TABLE 1. Comparison of the performances of different models

Method	In Sample MAE for San Juan	In Sample MAE for Iquitos	Out of Sample MAE
Linear Regression (Benchmark)	24.88	7.04	NA
Possion Regression	24.77	7.05	NA
Negative Binomial Regression	**24.76**	**6.98**	**26.14**
Random Forest Regression with Weight	**23.19**	**7.64**	**19.93**

The validation results are listed in Table 1. The benchmark Linear Regression was able to reach a mean absolute error of 24.88 for San Juan and 7.04 for Iquitos. The Poisson Regression was able to reach a mean absolute error of 24.77 for San Juan and 7.05 for Iquitos, surprisingly, didn't improve a lot compare to linear regression. The Negative Binomial Regression was able to reach a mean absolute error of 24.76 for San Juan and 6.98 for Iquitos, which appears to be the best model among the regression models. On the other hand, the machine learning Random Forest Regression model was able to reach a mean absolute error of 23.19 for San Juan and 7.64 for Iquitos, which ranks the best among all models.

Since Negative Binomial Regression and Random Forest Regression achieve the best results, I used the Negative Binomial Regression and Random Forest Regression to produce the submission file with the actual test file and submitted them into the competition cite. The Negative Binomial Regression's mean absolute error increased to 26.137, ranked 286th place among 11046 competitors; the Random Forest Regression with weight assigning was able to reach a mean absolute error of 19.93, which boost my rank from 286 to 250.

VII. Discussion & Future Work

Based on the performance of the Negative Binomial Regression, it is obvious to conclude that those regression models are overfitting the datasets. One possible reason might be that the correlation between each feature and the total_cases was less. Therefore the relationship between those features and total_cases may be non-linear. Another possible reason might be the correlation between selected features since they are unclear.

This may also be the reason why Random Forest Regression models perform ahead of other models due to the ability to learn non-linear relationships and robustness to outliers since it uses multiple models to obtain better predictive performance by using ensembling methods and cross-validation to select the best number of trees.

To sum up the project, I first fill the missing values with linear interpolation and apply z-score normalization in the preprocessing step. When it comes to feature selection, I first analyzed the correlation matrix containing correlation between each feature and remove the high correlated ones, and then I cross-compare with the correlation between features and the total cases and feature's seasonal decompose to select features with high correlation and significant seasonal periodic patterns. For Random Forest Regression, I also added weights to each feature. In the end, I was able to achieve MAE 19.93 and reached the top 25th quantile of the competition.

However, the project did not use the “weekofyear” feature. By observation, there should be a connection between this variable and the total cases. I believe one of the possible approaches is to use “weekofyear” to predict the seasonal cases first, then use the seasonal cases to predict the total cases. What’s more, another approach is to use a stack model, which produces San Juan and Iquitos with different models, and combine the results together.